FinFET-Based Power Simulator for Interconnection Networks

CHUN-YI LEE and NIRAJ K. JHA
Princeton University

Double-gate FETs, specifically FinFETs, are emerging as promising substitutes for bulk CMOS at the 32nm technology node and beyond because of the various obstacles to scaling faced by CMOS, such as short-channel effects, leakage power, and process variations. Another trend in chip multiprocessor design is incorporation of sophisticated on-chip interconnection networks. However, such networks are significant power-consumers. In this article, we address these two trends by presenting a power simulator for FinFET-based on-chip interconnection networks. It estimates both dynamic and leakage power. We present results for various FinFET design styles and temperatures (since leakage power changes drastically with temperature), and show that one FinFET design style may be much superior to another from the power consumption point of view.

Categories and Subject Descriptors: B.7.1 [**Integrated Circuits**]: Types and Design Styles—*Advanced technologies*; B.3.3 [**Memory Structures**]: Performance Analysis and Design Aids—*Simulation*; C.1.2 [**Processor Architectures**]: Multiple Data Stream Architectures (Multiprocessors)—*Interconnection architectures*

General Terms: Design, Performance

Additional Key Words and Phrases: FinFETs, interconnection network, power consumption, power simulator

ACM Reference Format:

Lee, C.-Y. and Jha, N. K. 2010. FinFET-Based power simulator for interconnection networks. ACM J. Emerg. Technol. Comput. Syst. 6, 1, Article 2 (March 2010), 18 pages.
DOI = 10.1145/1721650.1721652 http://doi.acm.org/10.1145/1721650.1721652

1. INTRODUCTION

With increasing wire delay and demand for high bandwidth [ITRS 2007; Dally and Towles 2001; Ho et al. 2001], packet-switched on-chip interconnection networks have been attracting increasing attention recently. On-chip interconnection networks connect processor cores and memories through on-chip routers

This work was supported by SRC under Contract No. 2007-HJ-1602 and 2008-HJ-1793, and NSF under Grant No. CNS-0613074.
Authors' addresses: C.-Y. Lee (corresponding author), N. K. Jha, Department of Electrical Engineering, Princeton University, Princeton, NJ 08544; email: {chunlee, jha}@princeton.edu.

DOI 10.1145/1721650.1721652 http://doi.acm.org/10.1145/1721650.1721652

and links. A structured on-chip network not only provides enhanced scalability, but also enables the use of shorter wires, which reduce the packet traversal latency. Due to the availability of sufficient on-chip wires, interprocessor links are not constrained by chip pin counts. Hence, high bandwidth as well as performance can be achieved.

To explore and verify the performance of on-chip interconnection networks, a number of Chip MultiProcessor (CMP) architectures have been proposed and fabricated, both in academia [Taylor et al. 2004; Sankaralingam et al. 2003; Mai et al. 2000] and industry [Pham et al. 2006; Mukherjee et al. 2002; Hoskote et al. 2007]. As the number of processor cores in a CMP increases, the power consumption of interconnection networks is becoming a critically important issue [Wang et al. 2002; Chen and Peh 2003; Shang et al. 2003]. It is reported that in the Alpha 21364 processor [Mukherjee et al. 2002], the integrated router and links consume 25W, which is about 20% of its total power of 125W. In Intel's teraflops processor [Hoskote et al. 2007], the power consumption of its router plus links is about 28% of the tile power, when operated at 4GHz. The InfiniBand switch dissipates up to 37.5% of the server blade's power budget of 40W [Wang et al. 2002]. Moreover, as bulk CMOS scales down to the 32nm technology node, it is facing various obstacles, such as Short-Channel Effects (SCEs), process variations, dopant fluctuations, increased leakage power, etc. Although system-level techniques have been proposed [Chen and Peh 2003] to reduce leakage power, the underlying transistor-level solutions are still urgently needed to overcome the leakage problem.

FinFETs have emerged as promising devices for implementing nanoscale circuits due to their better scalability compared to traditional bulk CMOS devices [ITRS 2007]. In addition, FinFETs offer the following four major advantages: (1) less leakage current compared to traditional CMOS transistors, (2) independent back-gate bias control to dynamically manage leakage power consumption, (3) less area because fin height determines effective channel width, and (4) more resistance to process variations due to the use of a lightly doped channel.

In this work, a FinFET-based power simulator, called ORION-FinFET, is discussed that can enable users to explore network power consumption under various FinFET design styles and temperatures. The simulator is built on top of the popular ORION [Wang et al. 2002] infrastructure that is targeted at bulk CMOS. FinFET power models and a power library are presented. The power models describe how the power of router components is calculated from the capacitance and leakage current of logic gates. The power library specifies the capacitance and leakage current values of logic gates for various FinFET operation modes and at different temperatures. The original power models in ORION are revised for FinFET-based router components. New power models are added to estimate the power consumption of the register file, clock tree, and interrouter links. Moreover, different FinFET power libraries are included for various FinFET technology nodes (45nm, 32nm, 25nm). In summary, the power simulator will enable computer architects to build routers with different FinFET operation modes and explore the power breakdown among its various components at an early design stage. Architects may also explore the use of

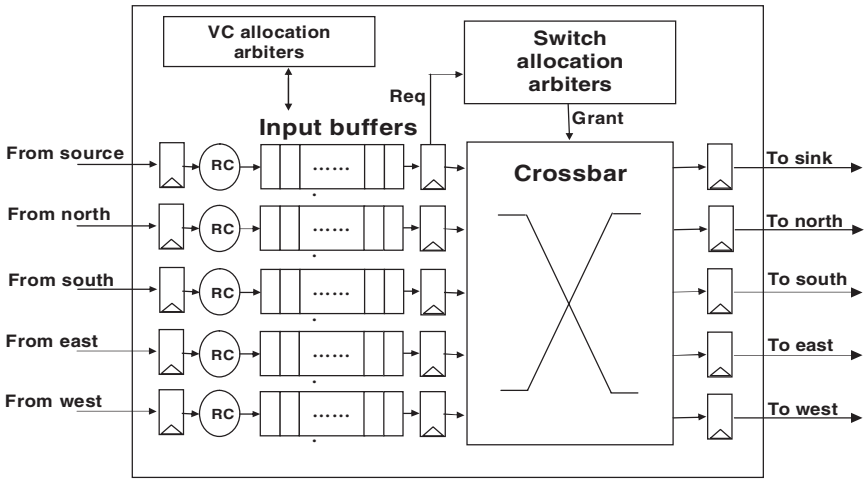


Fig. 1. Router microarchiteture.

different FinFET modes for different router components to further optimize router performance and power consumption.

The remainder of this article is organized as follows. Section 2 introduces background material and previous work related to this topic. Section 3 discusses our implementation methodology in detail. Section 4 explores the power breakdown of FinFET-based interconnection networks. Section 5 concludes.

2. BACKGROUND

In this section, background material is introduced. The on-chip router microarchitecture is discussed first. Such a router is a fundamental building block of an interconnection network. ORION [Wang et al. 2002], a previous power simulator for interconnection networks, is then discussed. Then, to understand how the proposed FinFET-based power simulator works, the FinFET structure and its operation modes are briefly described.

2.1 On-Chip Router Microarchitecture and ORION

Figure 1 shows the microarchitecture of an on-chip router. The function of a router is to route a packet from its input port to its destination port. There are five input/output ports in a typical router, assuming that the network topology is the popular two-dimensional mesh network. Four ports correspond to the four neighboring directions, whereas the fifth is used as the router's local access port. There are three major components that constitute an on-chip router: buffers, crossbar, and arbiters. They are the major power consumers.

Figure 2 shows the pipeline stages of a modern router. The pipeline stages are traversed by a flit, which is the unit of a packet that can be routed at a time. Usually, the flit width equals the bit width of the input ports and links. As a flit arrives at the head of the input buffer, a Route Calculation (RC) unit determines an output port for the flit. If the flit is the header of the

Fig. 2. Router pipeline stages.

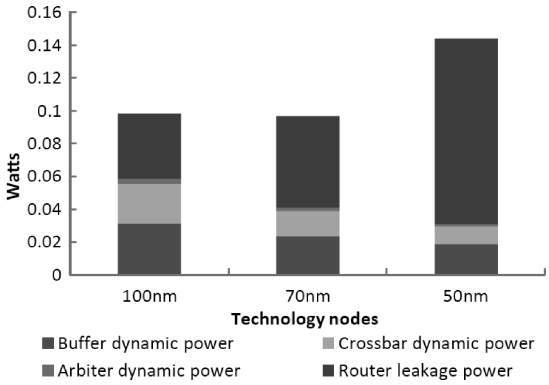


Fig. 3. Router power breakdown for bulk CMOS.

packet, it undergoes arbitration for Virtual Channels (VCs). This corresponds to the VC allocation stage. Next, in the Switch Allocation (SA) stage, the arbiters determine which flits have the right to access the inputs of the crossbar. The winning flit traverses the crossbar in the Switch Traversal (ST) stage. Finally, the flit leaves the router, and traverses the link to the downstream router. This corresponds to the Link Traversal (LT) stage. To model the power of routers in an interconnection network, all of these stages have to be carefully considered.

ORION is a power-performance simulator for on-chip bulk CMOS routers [Wang et al. 2002]. It estimates router power consumption using built-in dynamic and leakage power models for buffers, crossbars, and arbiters [Wang et al. 2002; Chen and Peh 2003]. Its power models have been validated against the MIT RAW processor [Taylor et al. 2004], with power estimation error less than 10%. However, ORION does not incorporate clock and link power in its simulations. In Hoskote et al. [2007], it is estimated that when operated at 4GHz, clock and link power may constitute about 50% of tile power, and thus cannot be ignored. In addition, ORION only estimates leakage power at a fixed temperature (80°C) [Chen and Peh 2003]. As a result, to build a FinFET-based network power simulator for deep submicron technology nodes, several further revisions of ORION are necessary.

Figure 3 compares the router power breakdown reported by ORION for three different technology nodes. The operating frequency is 1GHz. It can be seen that leakage power grows significantly as technology scales down. At the 50nm node, leakage power is 2.8X relative to the 100nm node. This is mainly attributable to SCEs. ORION calculates the power breakdown at these technology nodes by scaling from the 100nm node. Therefore, exact transistor sizes for building the

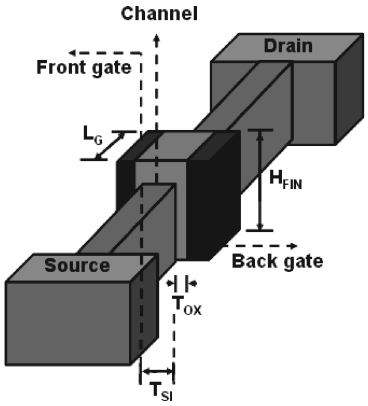


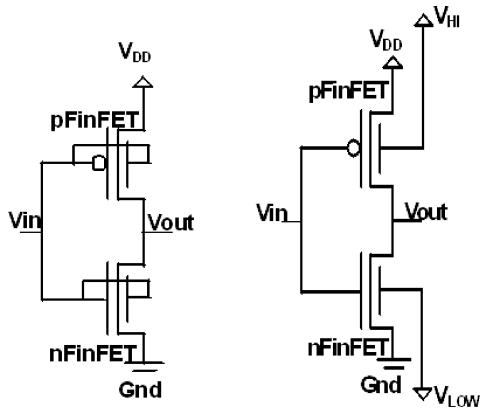
Fig. 4. FinFET structure.

router components are not used. In our work, FinFET as well as logic gate sizes are carefully calibrated at each technology node, and not simply scaled.

2.2 FinFET Structure and SG/LP Modes

The FinFET structure is shown in Figure 4. It uses a fin as the channel body. The gate length (L_G) is equal to the fin length. The gate width is quantized by the number of fins, and is obtained by multiplying the fin count with twice the fin height (H_{FIN}). The front and back gates are on opposite sides of the fin, and can either be shorted or independently biased to different voltages. Fin thickness (T_{SI}) and gate-oxide thickness (T_{OX}) greatly affect the amount of leakage current. FinFETs have been proven to provide relatively lower leakage current than traditional bulk CMOS FETs. As CMOS devices scale down to submicron nodes, thinner gate-oxide and higher channel/body doping densities are required to overcome SCEs. A thinner gate-oxide induces more gate-tunneling current, and the resulting leakage power consumption becomes intolerable beyond a certain point. In addition, a higher channel/body doping concentration degrades subthreshold swing. With a narrow fin and electrically coupled double gates, FinFETs can effectively suppress SCEs without utilizing a very thin gate-oxide. Heavy channel doping is not required for SCE control in FinFET devices. Thus, process variations due to random dopant fluctuation effects can be ameliorated.

Shorted-Gate (SG) and Low-Power (LP) modes are two possible ways to implement FinFET logic gates [Muttreja et al. 2007]. Figure 5 illustrates SG and LP mode inverters. In the SG mode, the front and back gates are tied together. In the LP mode, the back gate of the pFinFET (nFinFET) is reverse-biased to a separate V_{HI} (V_{LOW}). The threshold voltage of the front gate varies linearly with the back-gate bias voltage, and can be adjusted to the desired value [Muttreja et al. 2007]. Hence, the leakage current of an LP mode gate can be significantly reduced by reverse biasing the back gate. The disadvantage of an LP mode gate is increased gate delay due to the higher threshold voltage. This can be overcome by increasing the fin count of the LP mode FinFETs to increase



(a) SG mode (b) LP mode

Fig. 5. FinFET modes.

Table I. Capacitanc, Leakage Current, Delay, and Power for the Inverters in Figure 5

Mode	V_H/V_{LOW} (V)	Cg (aF)	Cd (aF)	I_{off} (nA)	FO4 Delay (ps)
SG	N/A	60.0	21.8	26.19	2.90
LP	1.2/−0.2	33.6	22.2	1.80	4.67
	1.3/−0.3	34.6	22.4	0.56	5.23
	1.4/−0.4	34.8	22.6	0.17	5.98

its on-state current. This induces little area overhead because of the vertical nature of fins and small inter-fin spacing. Also, although an increased fin count may induce slightly higher dynamic power, the reduction in leakage power far outweighs it.

Table I shows back-gate bias values (V_{HI}/V_{LOW}), gate/drain capacitances (Cg/Cd), off-state current consumption (I_{off}), and the fanout-of-four (FO4) delay for the inverters in Figure 5 at the 32nm technology node, where V_{DD} is assumed to be 1V. The numbers are simulated using UFDG [Fossum et al. 2004], which is an accurate process/physics-based double-gate MOSFET model. The temperature in this instance is set to 105°C. The fin count is set to 2 for the pFinFET and 1 for the nFinFET. It can be seen that the gate capacitance of the LP mode inverter is only half that of the SG mode inverter. This is because at the same fin count, an SG mode inverter controls both the front and back gates. Relative to the SG mode, the LP mode enables I_{off} to be reduced by an order of magnitude. However, when the back-gate bias voltages of the LP mode inverter are set to $V_{HI} = 1.4V$ and $V_{LOW} = -0.4V$, the FO4 delay of the LP mode inverter is approximately 2X that of the SG mode inverter. As a result, there is a trade-off between delay and I_{off} when choosing logic gates.

3. IMPLEMENTATION METHODOLOGY

In this section, the power simulation flow of the FinFET-based interconnection network is introduced. The network is divided into three components for

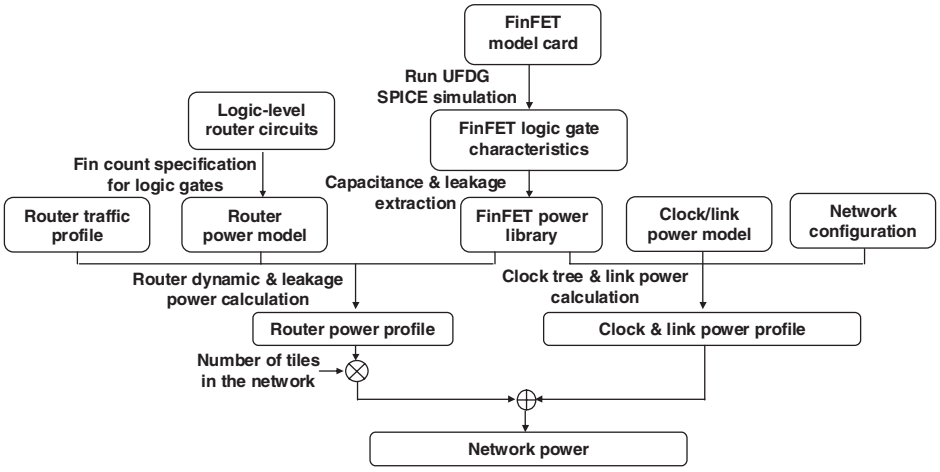


Fig. 6. Interconnection network power simulation flow.

power estimation: routers, clock tree, and links. The power of each component is estimated by using its parameterized power model. The power model calculates dynamic and leakage power based on our FinFET power library. This library specifies the capacitance and leakage current of minimum-size logic gates, which may be implemented in either SG or LP mode, depending on the simulation requirement.

3.1 Simulation Flow

Figure 6 illustrates the overall simulation flow of our FinFET-based network power simulator. The simulator core comprises five major parts: router traffic profile, router power model, FinFET power library, clock/link power model, and network configuration. The first three generate the router power profile, while the last two and the FinFET power library generate the clock/link power profile. The router power profile provides the power breakdown of the router components in a single router. The clock/link power profile provides the power breakdown of the clock tree and inter-router links of the entire network. The total power consumed by the routers is obtained by multiplying the power of a single router with the number of tiles in the network. The power of the entire network is obtained by combining the power of routers with that of the clock tree and links.

The router traffic profile and network configuration are provided by the user. The router traffic profile specifies the input traffic load of each router port. A heavier traffic load leads to a higher dynamic power. In this work, if not specifically stated, the traffic load of all router ports are simulated by GARNET [Agarwal et al. 2009], assuming uniform random traffic. Uniform random traffic assumes that packets are uniformly injected into the network at each router, with randomly distributed destinations. A traffic load of 0.1 implies that packets are received at each router port at the rate of 0.1 of the clock frequency. Other

Table II. 32nm FinFET Design Parameters

Parameters	FinFET
L_G (nm)	13
H_{FIN} (nm)	30
T_{si} (nm)	7.0
T_{ox} (nm)	1.0
Channel doping, N_{BODY} (cm^{-3})	10^{15}
V_{DD} (V)	1.0
R_{SD} ($\Omega - \mu$)	170
Source/drain doping, N_{DS} (cm^{-3})	10^{20}

traffic loads and patterns can also be characterized using the same simulation flow.

The network configuration specifies the size of the tiles, link length, flit width, and network topology. The link length and tile size determine the required wire length. The longer the wires, the higher the wire capacitance they contribute, which directly affects the dynamic power consumed by the clock tree and links. The network topology determines the number of routers in the network, and how they are connected by the links.

3.2 FinFET Power Library

A complete FinFET power library consists of a capacitance library and a leakage current library for the SG and LP mode gates at various temperatures. The capacitance library contains the wire capacitance per unit length and the gate/drain capacitance of logic gates used in the circuit. The types of logic gates include NAND, NOR, and inverter. AND, OR, XOR, buffer, and any other type of logic gate are formed from these primitive gates. Only the values for minimum-size gates are included, because the gate sizes are actually quantized by fin counts. Values for larger gates can be linearly scaled from those of the minimum-size gates. The wire capacitance values are obtained from ITRS [2007], which includes capacitance values per unit length for local, intermediate, and global wires. Local wires are used to connect logic gates, intermediate wires serve as the communication link between router components, and global wires are mainly adopted for the clock tree and links between routers.

The leakage current library contains the leakage current for the minimum-size gates for temperatures ranging from 20°C to 120°C. For the LP mode, this library also includes the leakage current under various back-gate bias voltages. Since leakage current is proportional to the fin count, the leakage current of larger gates can be linearly scaled from that of minimum-size logic gates.

The UFDG SPICE model [Fossum et al. 2004] is used to simulate 32nm FinFET logic gates (other technology nodes are discussed later). Minimum-size logic gates (NAND, NOR, and inverters) are simulated for their gate/drain capacitance, leakage current, and FO4 delay. The key transistor-level design parameters are summarized in Table II. The values of L_G, T_{OX}, and R_{SD} for the 32nm technology node are obtained from ITRS [2007]. N_{BODY} and N_{DS} are typical values suggested by UFDG. T_{SI} is set to 7.0nm, which is within the typical range of $0.5L_G \sim 0.7L_G$.

Gate/drain capacitances and leakage current are simulated for logic gates in order to evaluate the dynamic and leakage power. The dynamic power is consumed by charging and discharging of the capacitors. It can be calculated by the equation

$$P_{Dynamic} = 0.5 \times \alpha \times f_{Clk} \times C_{Load} \times V_{DD}^2, \quad (1)$$

where α is the switching activity, f_{Clk} the clock frequency, V_{DD} the supply voltage, and C_{Load} the load capacitance of the logic gate. Leakage power, which depends heavily on the working temperature, can be calculated as

$$P_{Leakage} = \Sigma_s Prob(s) \times I_{Leak}(s, T) \times N_{Fin} \times V_{DD}, \quad (2)$$

where Prob(s) is the probability of being in input state s, T the current temperature, N_{Fin} the number of fins, and $I_{Leak}(s,T)$ the leakage current per fin at input state s and temperature T.

3.3 Power Models

The building blocks of an interconnection network that we model fall into the following three categories.

(1) On-chip router: buffers, crossbar, arbiters, and D flip-flop (DFF) arrays.
(2) Clock tree: Local and global clock trees, drivers, clock wires, and loads.
(3) Interrouter links: link wires and drivers.

For each category, the dynamic and leakage power of its components are modeled in a hierarchical fashion. This hierarchy traverses the highest architectural level to the lowest FinFET level. The power of any component is derived from its lower-level building blocks. For example, for the SRAM input buffer, the power of the wordline, bitline, SRAM cells, sense amplifier, and precharge circuitry have to be calculated first. At the lowest FinFET level, the power of a minimum-size FinFET logic gate is directly derived from the FinFET power library.

The FinFET-based router power model is revised from ORION's model for bulk CMOS. ORION's power model includes capacitance equations [Wang et al. 2002] and leakage current tables [Chen and Peh 2003] for various router components. The capacitance equations formulate the load capacitance at each node of the circuit for calculating the dynamic power. The leakage current table specifies the leakage current per transistor width over length, which is similar to our leakage current library. In this work, the leakage current table is obtained by extending the leakage current library to specify the leakage current for various logic gates that were used in our simulator. For instance, this table specifies the leakage current of a simple inverter from the small one used in an SRAM cell to the large one used in a link driver. Given the fin count, the capacitance and leakage current of logic gates are linearly scaled from those of the minimum-size logic gates. The calculated capacitance values can be directly used with ORION's capacitance equations.

Table III shows an example power model for an SRAM buffer. This is borrowed from ORION's power model [Wang et al. 2002]. An SRAM buffer is

Table III. ORION power Model for an SRAM buffer [Wang et al. 2002]

Parameters	
B	Buffer size in flits
F	Flit size in bits
P_r	Number of buffer read ports
P_w	Number of buffer write ports
H_{cell}	Memory cell height
W_{cell}	Memory cell width
D_w	Wire spacing
T_p	Pass transistor connecting bitlines and memory cells
T_{wd}	Wordline driver
T_{bd}	Write bitline driver,
T_c	Read bitline precharge transistor
T_m	Memory cell inverter
Capacitance equations	
Wordline width	$L_{wl} = F(W_{cell} + 2(P_r + P_w)D_w)$
Bitline width	$L_{bl} = B(H_{cell} + (P_r + P_w)D_w)$
Wordline Cap.	$C_{wl} = 2FC_g(T_p) + C_a(T_{wd}) + C_w(L_{wl})$
Read bitline Cap.	$C_{br} = BC_d(T_p) + C_d(T_c) + C_w(L_{bl})$
Write bitline Cap.	$C_{bw} = BC_d(T_p) + C_d(T_{bd}) + C_w(L_{bl})$
Precharge Cap.	$C_{chg} = C_g(T_c)$
Memory cell Cap.	$C_{cell} = 2(P_r + P_w)C_d(T_p) + 2C_a(T_m)$
Sense amp. energy	Obtained from SPICE simulations

(*) Cg is the gate capacitance, C_d the drain capacitance, C_w the wire capacitance per unit length, C_a equal to $(C_g + C_d)$

first subdivided into wordlines, bitlines, SRAM cells, precharge transistors, and sense amplifiers. Capacitance equations for the first four components are derived based on the design parameters of the SRAM buffer. The energy consumed by the sense amplifier is obtained from the UFDG SPICE simulations. The dynamic power of each component can be evaluated according to the corresponding capacitance equations, switching activity, V_{DD}, and the operation frequency. The total power of the SRAM buffer is obtained by summing up the dynamic and leakage power of all its components. Power models for crossbars and arbiters can be found in Wang et al. [2002].

The input buffers can be implemented with either SRAM cells or register file cells [Kumar et al. 2007], depending on the desired application and performance. The length of the wordline and bitline, and the size of the 32nm FinFET-based SRAM cell are scaled from the layout values in Guo et al. [2005]. The register file cell module is a new feature in our simulator. It uses only one bitline, thus consumes less dynamic power than an SRAM cell. However, due to the fact that a register file cell has more transistors, it results in a larger area than that of an SRAM cell.

DFF arrays are inserted into the router at the following locations: before the input buffer, after the input buffer, and after the crossbar. The width of a DFF array is equal to the flit width. These arrays dissipate both dynamic and leakage power, and their capacitance contributes C_{Load} to the clock power Eq. (3). For dynamic power, DFF arrays share the same traffic load value as the input buffer. The power dissipated by DFF arrays is incorporated into buffer power consumption.

Clock/link power models are not provided in ORION, and are derived from Liu and Svensson [1994] and Chen et al. [1999]. An H-tree is adopted as the clock tree structure (Intel's teraflops processor chip [Hoskote et al. 2007] also employs it). Clock drivers are distributed across the entire H-tree, rather than placing a single large driver at the root. The distributed driver scheme is much more power efficient than the single driver scheme [Chen et al. 1999]. Clock drivers are inserted at each branch root of the clock tree. Driver sizes are designed such that the delay of each branch is equal to an inverter FO4 delay. The clock power of a router can be calculated as

$$P_{Clk} = f_{Clk} \times (C_{H\text{-}tree} + C_{Load} + C_{Driver}) \times V_{DD}^2, \quad (3)$$

where $C_{H\text{-}tree}$ is the wire capacitance of the H-tree, C_{Load} the total load capacitance, and C_{Driver} the sum of the total driver capacitance in the H-tree. $C_{H\text{-}Tree}$ can be computed as

$$C_{Wire} = C_U \times D_R \times [8 \times 1/4 + 4 \times 1/4 + 2 \times 1/2 + 1 \times 1/2], \quad (4)$$

where C_U is the global wire capacitance per unit length, D_R = sqrt(A_R), and A_R is the area of the router.

The link power depends on the flit width and the number of links in the network. It can be calculated as

$$P_{Link} = 0.5 \times \alpha \times f_{Clk} \times (C_U \times L + C_{Load} + C_{Driver}) \times V_{DD}^2 \times F \times N_{Link}, \quad (5)$$

where F is the flit width, N_{Link} the total number of the links in the network, and L the length of a single link. Note that the leakage power dissipated by the clock net and links is only contributed by the drivers in them.

The whole network power is obtained by summing the router power, clock power, and link power, which can be summarized using the equation

$$P_{Network} = N_{Tile} \times (P_{Buffer} + P_{Crossbar} + P_{Arbiter}) + P_{Link} + P_{Clk} + P_{Tleak}, \quad (6)$$

where N_{Tile} is the total number of tiles in the network, and P_{Tleak} the total leakage power in the interconnection network.

4. EXPERIMENTAL RESULTS

In this section, experimental results are provided to compare and evaluate the performance of FinFET-based interconnection networks. The power breakdown of the network components is explored first for both SG and LP modes. The network power breakdowns for bulk CMOS and FinFETs are also presented. The network latency and power at different packet injection rates are explored. The comparison of network power consumption for different traffic patterns is plotted. The trend for leakage power versus temperature is illustrated. Finally, the power breakdown for different technology nodes is provided.

4.1 Simulation Setup

For illustration purposes, we assume a 4×4 mesh network. Each router has 5 ports, with 48 flit buffers per port. A 5×5 crossbar is used. The flit width is set to 128 bits. Each packet consists of four flits. If not specifically stated, uniform

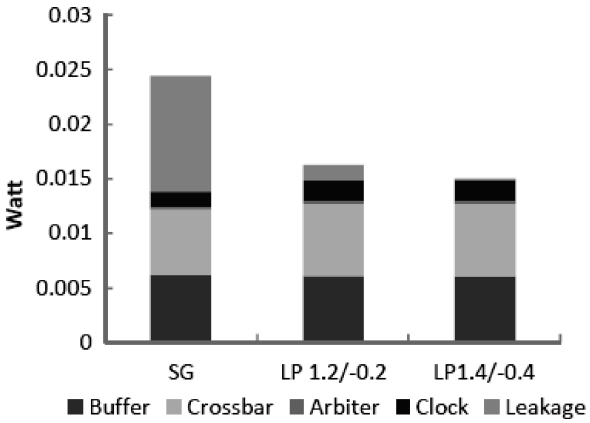


Fig. 7. Router power breakdown for SG/LP modes.

random traffic is assumed for the entire network, with the traffic injection rate of each router set to 0.1 packets/node/cycle. The clock frequency is set to 1GHz. The distance between two tiles is assumed to be 1mm. The operating temperature is assumed to be 105°C if not specifically stated. The fin counts of the LP-mode gates are assumed to be 2X that of SG-mode gates in order to compensate for the delay degradation.

4.2 Power Breakdown Comparison

Figure 7 compares the power breakdown for the SG- and LP-mode routers. Two sets of back-gate bias voltages are used for the latter. Dynamic and leakage power are presented separately for a clear comparison. Dynamic power includes contributions from buffers, crossbar, arbiters, and local clock tree. Only the local clock tree within the router is included in this power breakdown, because the global clock distribution net should not be counted as part of the router power. As can be seen, the dynamic power of the LP-mode routers is slightly higher than that of the SG-mode router. This is due to the fact that twice as many fins are used in the LP-mode routers. However, leakage power is significantly reduced due to LP-mode FinFETs. When the back-gate bias voltage is set to $V_{HI} = 1.4V$ and $V_{LOW} = -0.4V$, the leakage power of the LP-mode router becomes very small, reducing total power by 38.5%.

Figure 8 shows the power breakdown of the entire network for both the SG and LP modes. In this figure, the router power includes all the dynamic and leakage power dissipated by the routers in the network. The clock power comes from the clock drivers and clock wires in the global clock distribution net. The link power comes from the traversal of packets over the inter-router links. The leakage power in this figure is solely contributed by clock and link drivers. An LP-mode router with back-gate bias voltage of $V_{HI} = 1.4V$ and $V_{LOW} = -0.4V$ consumes 31.1% less power than an SG-mode router.

Figure 9 compares the dynamic and leakage power for input buffers built from SRAM cells and register file cells. The FinFETs are assumed to be in the SG mode. As can be seen, the dynamic power of the register file is lower than

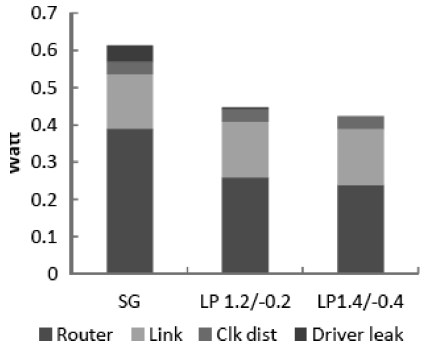


Fig. 8. Entire network power breakdown.

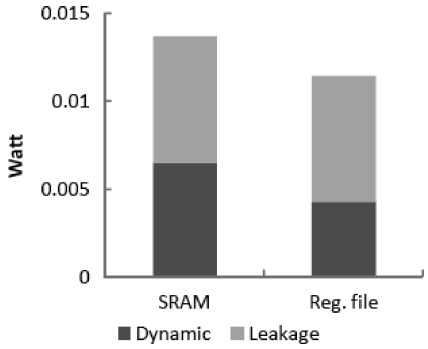


Fig. 9. SG-mode SRAM vs. register file power.

that of the SRAM-based buffer. This is attributable to the fact that each register file cell requires only one bitline, rather than two bitlines used by the SRAM cell. The leakage power of register file cells is comparable to that of SRAM cells even though it uses a few more transistors.

4.3 Bulk CMOS versus FinFETs

Figure 10 shows the comparison for the power breakdown of the entire network for both bulk CMOS and LP-mode FinFETs. The simulations are all performed using the latest 32nm predictive technology model (PTM) [PTM; Zhao and Cao 2006] rather than UFDG. This is because PTMs are available for bulk CMOS and FinFETs at the 32nm technology node, whereas UFDG only models FinFETs. The fin height provided by the 32nm PTM is equal to 40nm. The gate lengths of bulk CMOS and FinFETs are both set to 32nm, which is the default value assumed by the PTM. The gate widths of bulk CMOS are adjusted to be the same as those used for the LP-mode FinFETs. In this figure, the router power only includes the dynamic power dissipated by the routers. The leakage power comes from the routers, link drivers, and clock drivers in the network. As can be seen, the leakage power of a bulk CMOS network is 2.7X higher as compared to an LP-mode FinFET network with the back-gate bias voltage of V_{HI} =1.2V and $V_{LOW} = -0.2$V.

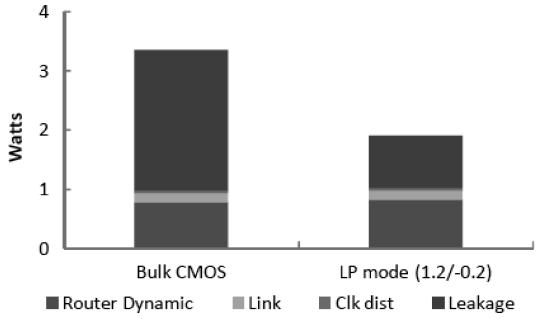


Fig. 10. Network power breakdown for bulk CMOS and FinFETs.

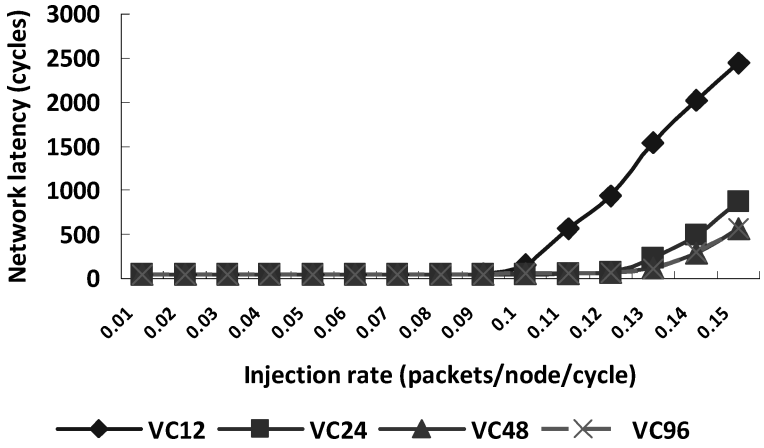


Fig. 11. Network latency vs. injection rate.

4.4 Latency and Power vs. Injection Rate

Figures 11 and 12 plot network latency and power as a function of packet injection rate, respectively. The traffic pattern applied is uniform random. Four router configurations are simulated and compared:

(1) 3 VCs per port and 4-flit input buffer per VC (VC12).
(2) 6 VCs per port and 4-flit input buffer per VC (VC24).
(3) 12 VCs per port and 4-flit input buffer per VC (VC48).
(4) 12 VCs per port and 8-flit input buffer per VC (VC96).

In Figure 11, only one line is plotted for each router configuration, because the LP-mode routers are designed such that their pipeline latency is the same as the SG-mode routers. It can be seen that the network latency grows exponentially as the packet injection rate increases. A high packet injection rate results in competition for router resources, and hence, packets have to wait in the router pipeline for more cycles. The contention becomes even worse when there are fewer VCs per port. When the injection rate is 0.15 packets/node/cycle, the network latency of VC12 is 4.3X as compared to VC48. VC96 does not

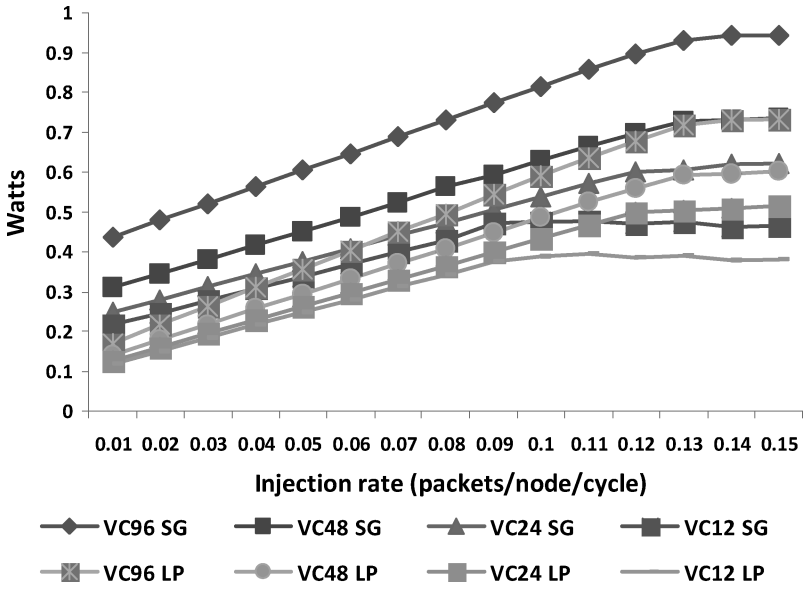


Fig. 12. Network power vs. injection rate.

improve network latency significantly as compared to VC48, because the input buffer size of its VC is larger than the packet size. Therefore, the extra buffer space is wasted. The VC96 case is shown here to demonstrate how our simulator allows computer architects to evaluate the performance of different router architectures.

Figure 12 compares the power consumption of the four router configurations. The network power is plotted for both SG- and LP-mode routers. The power consumption increases as the packet injection rate increases. This is attributable to increased switching activities caused by increased packets. SG-mode routers consume more network power than LP-mode routers for each router configuration. Comparing Figure 11 with Figure 12, it can be seen that VC96 consumes more power than VC48 with little improvement in network latency. Therefore, VC48 is preferable over VC96 with these simulation settings.

4.5 Power Comparison for Various Traffic Patterns

Figure 13 compares the power consumption for both SG- and LP-mode routers for three kinds of traffic patterns. These traffic patterns include random, bit-inverse, and tornado traffic, and were simulated by GARNET [Agarwal et al. 2009]. For the random traffic pattern, packets are sent to destinations in a uniform random fashion. The bit-inverse traffic sends packets from nodes to their bit-inverse nodes. For tornado traffic, packets are passed from router to router in a tornado-like manner. Different traffic patterns have different link and router port utilization rates, thus dissipate different amounts of dynamic power. One can see that the power reduction due to the use of the LP mode, with back-gate bias voltage of $V_{HI} = 1.2V$ and $V_{LOW} = -0.2V$, is similar for all

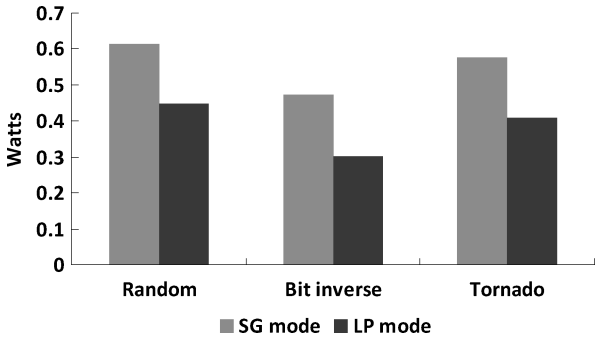


Fig. 13. SG- and LP-mode (1.2V/−0.2V) network power comparison for different traffic patterns.

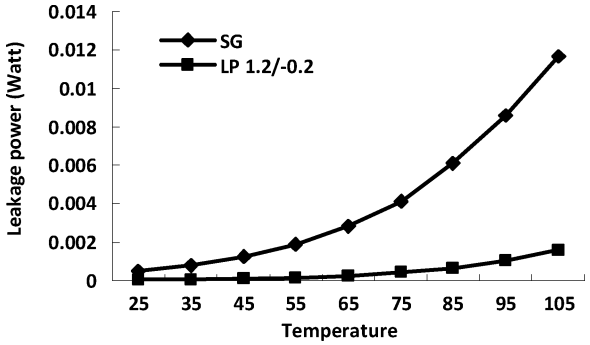


Fig. 14. Router leakage power vs. temperature.

the traffic patterns. This is mainly due to the fact that leakage power is not affected by network traffic. Therefore, power reduction with the LP mode is more significant when the network has lighter traffic.

4.6 Leakage Power versus Temperature

Figure 14 plots leakage power for the two routers as a function of temperature. The back-gate bias voltages of the LP-mode router are assumed to be $V_{HI} = 1.2V$ and $V_{LOW} = -0.2V$. The leakage power of the SG-mode router grows much faster than that of the LP-mode router as the temperature increases. At 105°C, the leakage power of the LP-mode router as compared to the SG-mode router is 13.5%. This is mainly attributable to the fact that reverse biased back gates lead to a higher threshold voltage, thus reducing leakage current. The leakage power of the LP-mode router with a back-gate bias of $V_{HI} = 1.4V$ and $V_{LOW} = -0.4V$ is not drawn because it is almost zero even at 105°C.

4.7 Power Comparison for Various Technology Nodes

Figure 15 shows the power breakdown comparison for LP-mode routers for three technology nodes. Again, the back-gate bias voltages of the LP-mode routers are assumed to be $V_{HI} = 1.2V$ and $V_{LOW} = -0.2V$. The fin count is kept the same for these three technology nodes, with the supply voltage fixed at 1V.

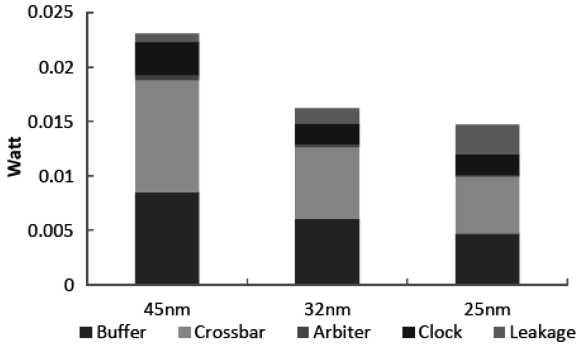


Fig. 15. LP-mode (1.2V/−0.2V) power breakdown for different technology modes.

As expected, total power decreases as the technology scales down. As compared to the 45nm technology node, the dynamic power of the router components (buffers, crossbars, and arbiters) at the 25nm node is lowered by 47.5%. This is mainly due to the reduction in gate size. The clock power only decreases by 38.2%, because the wire capacitance does not scale down so significantly [ITRS 2007]. At the 25nm node, leakage power increases to 3.6X, as compared to the 45nm node. Thus, as the technology scales, larger back-gate reverse biases or thinner fins will be required for better leakage current control.

5. CONCLUSIONS

In this article, a FinFET-based power simulator for interconnection networks was described. A power library for 32nm FinFET logic gates was used for the simulator. SG and LP modes were supported by the power library for various temperatures and back-gate reverse bias voltages. In addition to router power estimation, the power consumption of the clock tree, links, and DFF arrays were also estimated. The entire network power consumption can be obtained by summing the power of routers, clock tree, and links. Experimental results demonstrated that leakage power can be significantly reduced by using LP-mode FinFETs. The simulator provides a platform for computer architects to quickly estimate the power of FinFET-based interconnection networks at an early design stage.

REFERENCES

AGARWAL, N., PEH, L.-S., AND JHA, N. K. 2009. GARNET: A detailed interconnection network model inside a full-system simulator. In *Proceedings of the International Symposium on Performance Analysis of Systems and Software*. http://www.cs.wisc.edu/gems/download.html.

CHEN, R. Y., VIJAYKRISHNAN, N., AND IRWIN, M. J. 1999. Clock power issues in system-on-a-chip designs. In *Proceedings of the IEEE Computer Society Workshop on VLSI*. 48–53.

CHEN, X. AND PEH, L.-S. 2003. Leakage power modeling and optimization in interconnection networks. In *Proceedings of the International Symposium on Low Power Electronics Design*. 90–95.

DALLY, W. J. AND TOWLES, B. 2001. Route packets, not wires: On-Chip interconnection networks. In *Proceedings of the Design Automation Conference*. 684–689.

FOSSUM, J. G., GE., L., CHIANG, M.-H., TRIVEDI, V. P., CHOWDHURY, M. M., MATHEW, L., WORKMAN, G. O., AND NGUYEN, B.-Y. 2004. A process/physics-based compact model for nonclassical CMOS device and circuit design. *Solid State Electron. 48*, 6, 919–926.

GUO, Z., BALASUBRAMANIAN, S., ZLATANOVICI, R., KING, T.-J., AND NIKOLIC, B. 2005. FinFET-based SRAM design. In *Proceedings of the International Symposium on Low Power Electronics and Design*. 2–7.

HO, R., MAI, K., AND HOROWITZ, M. 2001. The future of wires. *Proc. IEEE 89*, 4, 490–504.

HOSKOTE, Y., VANGAL, S., SINGH, A., BORKAR, N., AND BORKAR, S. 2007. A 5-GHz mesh interconnect for a teraflops processor. *IEEE Micro 27*, 5, 51–61.

ITRS. 2007. International technology roadmap for semiconductors. http://www.itrs.net.

KUMAR, A.. KUNDU, P., SINGH, A. P., PEH, L.-S., AND JHA, N. K. 2007. A 4.6Tbits/s 3.6GHz single-cycle NoC router with a novel switch allocator in 65nm CMOS. In *Proceedings of the International Conference on Computer Design*. 63–70.

LIU, D. AND SVENSSON, C. 1994. Power consumption estimation in CMOS VLSI chips. *IEEE J. Solid-State Circ. 29*, 6, 663–670.

MAI, K., PAASKE, T., JAYASENA, N., HO, R., DALLY, W. J., AND HOROWITZ, M. 2000. Smart memories: A modular reconfigurable architecture. In *Proceedings of the International Symposium on Computer Architecture*. 161–171.

MUKHERJEE, S. S., BANNON, P., LANG, S., SPINK, A., AND WEBB, D. 2002. The alpha 21364 network architecture. *IEEE Micro 22,* 1, 26–35.

MUTTREJA, A., AGARWAL, N., AND JHA, N. K. 2007. CMOS logic design with independent-gate FinFETs. In *Proceedings of the International Conference on Computer Design.* 560–567.

PHAM, D. C., AIPPERSPACH, T., BOERSTLER, D., BOLLIGER, M., CHADHURY, R., COX, D., HARVEY, P., HARVEY, P. M., HOFSTEE, H. P., JOHNS, C., KAHLE, J., KAMEYAMA, A., KEATY, J., MASUBUCHI, Y., PHAM, M., PILLE, J., POSLUSZNY, S, RILEY, M., STASIAK, D. L., SUZUOKI, M., TAKAHASHI, O., WARNOCK, J., WEITZEL, S., WENDEL, D., AND YAZAWA, K. 2006. Overview of the architecture, circuit designs and physical implementation of the first generation cell processor. *IEEE J. Solid-State Circ. 41*, 1, 179–196.

PTM. Predictive technology model. http://www.eas.asu.edu/~ptm/.

SANKARALINGAM, K., NAGARAJAN, R., LIU, H., KIM, C., HUH, J., BURGER, D., KECKLER, S., AND MOORE, C. R. 2003. Exploiting ILP, TLP, and DLP with the polymorphous TRIPS architecture. In *Proceedings of the International Symposium on Computer Architecture*. 422–433.

SHANG, L., PEH, L.-S., AND JHA, N. K. 2003. Dynamic voltage scaling with links for power optimization of interconnection networks. In *Proceedings of the International Symposium on High-Performance Computer Architecture.* 91–102.

TAYLOR, M. B., LEE, W., MILLER, J., WENTZLAFF, D., BRATT, I., GREENWALD, B., HOFFMANN, H., JOHNSON, P., KIM, J., PSOTA, J., SARAF, A., SHNIDMAN, N., STRUMPEN, V., FRANK, M., AMARASINGHE, S., AND AGARWAL, A. 2004. Evaluation of the Raw microprocessor: An exposed-wire-delay architecture for ILP and streams. In *Proceedings of the International Symposium on Computer Architecture*. 2–13.

WANG, H., ZHU, X., PEH, L.-S., AND MALIK, S. 2002. ORION: A power-performance simulator for interconnection networks. In *Proceedings of the International Symposium on Microarchitecture*. 294–305.

ZHAO, W. AND CAO, Y. 2006. New generation of predictive technology model for sub-45nm early design exploration. *IEEE Trans. Electron. Dev. 53*, 11, 2816–2823.

Received February 2009; revised July 2009; accepted September 2009 by Siva Narendra