Microprocessors and Microsystems 39 (2015) 807–815

Contents lists available at ScienceDirect

Microprocessors and Microsystems

journal homepage: www.elsevier.com/locate/micpro

Thermal aware floorplanning incorporating temperature dependent wire delay estimation

Andreas Thor Winther[a], Wei Liu[b,*], Alberto Nannarelli[c], Sarma Vrudhula[d]

[a] Knowles Electronics, Roskilde, Denmark
[b] Oticon A/S, Smorum, Denmark
[c] DTU Compute, Technical University of Denmark, Kongens Lyngby, Denmark
[d] Computer Systems Engineering, Arizona State University, Tempe, USA

ARTICLE INFO

Article history:
Available online 23 October 2015

Keywords:
Temperature
Wire delay modeling
Floorplanning
Congestion
Reliability
Thermal analysis

ABSTRACT

Temperature has a negative impact on metal resistance and thus wire delay. In state-of-the-art VLSI circuits, large thermal gradients usually exist due to the uneven distribution of heat sources. The difference in wire temperature can lead to performance mismatch because wires of the same length can have different delay.

Traditional floorplanning algorithms use wirelength to estimate wire performance. In this work, we show that this does not always produce a design with the shortest delay and we propose a floorplanning algorithm taking into account temperature dependent wire delay as one metric in the evaluation of a floorplan. In addition, we consider other temperature dependent factors such as congestion and interconnect reliability.

The experiment results show that a shorter delay can be achieved using the proposed method.

1. Introduction

With technology scaling, the feature sizes of both CMOS devices and wires shrink and designers are able to integrate more and more functionalities into a single chip. The delay in CMOS transistors decreases as the channel length is reduced in each new process. The delay in metal wires, on the other hand, shows different behaviors. For local wires, the delay decreases as the distance between the end points becomes smaller with scaling. For global wires, which have to span across the chip, the delay increases due to the fact that the die size does not shrink but slightly increases in each new process [1]. In fact, the delay in global wires has increased steadily with technology scaling over the years and already dominates path delays [2].

In addition to technology scaling, the modeling of global wires is further complicated by thermal effects. Due to the high degree of integration (processing elements and memory blocks on the same die) and various aggressive power management techniques (such as clock gating, power gating, voltage islands, etc.), the power consumption in different regions of the chip (i.e. the power density) can vary significantly [3]. The spatially non uniform power consumption within the chip exhibits as thermal gradients, which are temperature differences between different regions.

The heat generated inside the chip has to be transferred to the ambient environment mainly through the heat sink attached to the silicon substrate. However, a secondary heat conduction path also exists from the substrate towards the packaging through the metal layers [4]. In nanometer technologies, in spite of an increase in the number of available metal layers, the top metal layers may still get closer to the substrate which results in a stronger thermal coupling between the substrate and the wires [5].

The high temperature and large thermal gradient in the metal layers can affect many aspects of interconnect design, including signal delay, routing congestion and reliability. The propagation delay in metal wires is severely degraded by high temperature as the electrical resistivity in metal increases linearly with temperature. The large within-die thermal gradients result in performance mismatch between wires of the same length but subject to different temperature. Traditional physical design algorithms, such as floorplanning and routing, assume resistivity in interconnects is uniform and constant. Consequently, wirelength is used as a metric to estimate signal delay and congestion of interconnects [6–8]. However, in designs where the substrate has a nonuniform thermal profile, the traditional way of estimating wire delay can lead to large errors. This is because wire performance decreases with an increase in temperature and the delay of a hot wire and a cool wire are no longer equal even though their lengths are the same.

Furthermore, the thermal effect is more significant in global wires than in local wires because global wires are routed in layers that are

[*] Corresponding author. Tel.: +45 60753957.
E-mail address: wli@oticon.dk, liuweizju@hotmail.com (W. Liu).

http://dx.doi.org/10.1016/j.micpro.2015.09.013
0141-9331/© 2015 Elsevier B.V. All rights reserved.

far away from the heat sink, and global wires span long distance thus possibly developing a larger thermal gradient.

Clock networks, which contain many global wires, are very sensitive to thermal variations. In recent years, temperature variation induced clock skew in clock distribution networks has received a lot of attention. In [9,10], the authors described design time clock tree synthesis algorithms to modify merging locations against nonuniform substrate thermal profile. In [11], optimal insertion of tunable delay buffers into clock trees is discussed to adjust at run time the delay of clock distribution paths that are more susceptible to temperature variations. Thermal aware global routing algorithms for improving reliability are also discussed in [12,13].

As for global signal wires, although extensive work has been done on thermal aware floorplanning, all of these works assume electrical resistivity in wires is constant and thermal gradients in the substrate have no impact on wire delay. These assumptions are in general invalid and increasingly inaccurate in nanometer high performance designs where large temperature gradients already exist in the substrate.

In this paper, we study the problem of estimating the temperature dependent wire delay during the floorplanning stage. We first illustrate the impact of nonuniform thermal profile on the delay in wires. Then we propose a new way to estimate the wire delay in thermal aware floorplanning algorithms. The proposed algorithm takes delay, instead of wirelength, as one of the optimization goals, in this way, mitigating the excessive delay overhead caused by high temperature. In addition, we also consider the impact of routing congestion and the reliability of wires, which are important metrics in evaluating floorplans in a realistic scenario.

2. Thermal aware floorplanning

Floorplanning is the initial stage of physical implementation of VLSI circuits, that determines, to a large extent, the quality of the final design. Floorplanning transforms the functional description of a circuit, in the form of a netlist of gates and macros, into a physical description, in the form of dimensions and location coordinates.

During the floorplanning stage, the main design tasks include macro block placement, global wire planning and Power/Ground network design. Traditional floorplanning algorithms only optimize the total area and wirelength. A smaller area reduces the cost since more circuits can be produced on the same wafer. Shorter wirelength makes routing easier and usually results in better performance as well.

In recent years, as thermal issues become prominent, the maximum temperature is also added to the cost functions in so called thermal aware floorplanning algorithms [8,14,15]. Placing a hot block in the middle of cool blocks can effectively bring down the peak temperature due to better heat spreading. Since the thermal coupling between high power consumption blocks can significantly affect the temperature distribution in the whole chip, thermal aware floorplanning algorithms consider peak temperature in addition to area and wirelength in the evaluation of a floorplan. The estimation of peak temperature usually requires the use of compact thermal models that can compute the temperature profile in a very efficient way [16].

The floorplanning tool proposed in [8], HotFloorplan, is an architectural level thermal aware floorplanner. HotFloorplan represents the topology of a floorplan in *Normalized Polish Expression* [6] and the optimization algorithm is implemented as a simulated annealing process. The algorithm chooses a random generated floorplan during each iteration of the annealing and uses the maximum temperature, total area and total wirelength as evaluation metrics in the cost function.

The pseudocode for the annealing is given in Algorithm 1. In brief, the optimization process goes through a series of steps and with every step a synthetic *temperature*[1] is changed according to an annealing schedule. The temperature is used to define how wide the search for a better floorplan is (within the solution space) – a low temperature means a narrow search.

Algorithm 1 Pseudocode for the HotFloorplan algorithm.

```
Set initial annealing temperature;
step = 0;
try = 0;
while steps != max steps AND probability > minimum do
  create initial floorplan;
  while try < max tries do
    try++;
    randomly perturb current floorplan;
    evaluate new floorplan;
    if new floorplan better than best floorplan then
      best floorplan = new floorplan;
    end if
    if accept(new floorplan) == true then
      current floorplan = new floorplan;
    end if
  end while
  step++;
  change simulated annealing temperature;
  calculate probability as a function of temperature;
end while
OUTPUT: Best floorplan
```

Within each step, HotFloorplan starts out with an initial floorplan that is simply any possible legal floorplan within the solution space. HotFloorplan then tries to optimize this floorplan by moving the blocks around. For every candidate floorplan, the algorithm invokes routines in HotSpot [17] to compute the worst case thermal profile using power dissipation values of each functional block. If the candidate has a lower cost, it is always accepted, otherwise the candidate is accepted conditionally based on a probability factor. At the end of the inner loop, we get the best known floorplan available at that temperature. The next step repeats this action only this time with another annealing temperature according to the annealing schedule. In the end, we get the best floorplan created during all steps. The algorithm either ends when the maximum number of steps has been reached, or if the probability (which is a function of the temperature) of the next step is under a given value (i.e., threshold). How HotFloorplan decides which floorplan is *the best* is vital, as it is this evaluation process that has been modified to include thermal awareness.

The connectivity information between blocks is stored in a two-dimensional connectivity matrix, and the wirelength between the two endpoints of a wire is estimated by measuring the “Manhattan distance”.

3. Delay modeling of metal wires

In this section, we describe the models used to estimate the temperature in wires and the process to calculate temperature dependent wire delay.

3.1. Temperature estimation

The temperature rise in metal wires is caused by both self-heating and heat diffusion from the substrate. During a signal transition inside a metal wire, the accelerated charge carriers (i.e., electrons) collide with other carriers and atoms in the electric field. The collisions

[1] Here, the term temperature is used as an attribute of the annealing and does not refer to the actual substrate (chip) temperature.