

Performance Improvement Using On-Board Wires for On-Chip Interconnects

Azad Naeemi, Payman Zarkesh-Ha, Chirag S. Patel, and James D. Meindl

Microelectronic Research Center, Georgia Institute of Technology, 791 Atlantic Dr. NW., Atlanta, Ga 30332-0269, USA
 Phone:(404) 894-9457 Fax: (404) 894-0462, Email: azad@ece.gatech.edu

Abstract

Utilizing a stochastic global-net length distribution for a projected GSI chip in year 2011, the number of total off-chip layers and pads required for a specified decrease of the maximum on-chip interconnect length are calculated. For example by adding four off-chip layers to the on-chip interconnects of the projected microprocessor, the global clock frequency can be increased from 3GHz to 4GHz, which is the maximum possible value limited by the time of flight delay.

Introduction

Global interconnects will pose a severe problem for future gigascale integrated systems (GSI). Rapid increase in the number of transistors increases the demand for wiring on-chip. Projected increase in the chip size and global clock frequency require low loss global interconnects to have an acceptable delay for long on-chip interconnects. An effective method of achieving low loss lines is to increase the cross-sectional area of the wires. However, this increase in the wiring cross-sectional area reduces the wiring density on the chip.

An alternate method of resolving the issue of lossy on-chip lines is to use off-chip wires available on the printed wiring board (PWB). Off-chip wires usually have large cross section and therefore even for the longest interconnects have negligible loss and if they are properly terminated, delay is determined by time of flight, which is a fundamental limit. However, the number of global interconnects that can be routed using off-chip wires on the PWB is limited by the minimum size line and space on the PWB. Therefore, a global intra-chip wiring model is required to determine an optimal distribution of global interconnects on-chip and off-chip.

New packaging technologies such as flip chip packages and high-density compliant wafer level packages provide very dense pad arrays with very small parasitics [1]. It is possible to use these packages and route some of the long interconnects through the PWB and take advantage of low loss off-chip wires.

Using a stochastic global wiring distribution for a gigascale chip it is shown that there are very few interconnects that are very long and by using a few off-chip wiring layers, maximum on-chip interconnect length can be reduced significantly. In this way, it is possible to increase the global clock frequency. Using a few PWB layers it is possible to reduce the maximum delay down to the time of flight of a corner-to-corner interconnect and hence increase the global clock frequency to the maximum possible value.

Extracting the Wiring Distribution

For extracting the wiring distribution, the International Technology Roadmap for Semiconductors (ITRS) projection for high performance microprocessors in year 2011 is used [2]. Key parameters are summarized in Table 1. Using a stochastic net-length distribution model for heterogeneous systems [3], the global net-length distribution of the projected chip is shown in Fig. 1. This figure shows number of nets having different lengths.

Each net consists of one driver and several receivers. The farthest receiver has the largest delay. If no repeater is used capacitance of the whole net contributes to the delay of the farthest receiver. However, since many repeaters are often inserted along the net; the path between the driver and the farthest receiver is isolated from capacitance of the branches by the repeaters at the beginning of each branch. Therefore, the length of the wire connecting the driver to the furthest receiver determines the largest delay of each net. Calling this length effective net length it is necessary to find the total net length versus effective net length. Assuming that nets are in the form of a tree shown in Fig. 2 segment length of each net can be found by

$$l_{seg} = \frac{l_{net}}{2(f.o.)} \quad (1)$$

Table 1. ITRS projections and some assumptions for a typical high performance microprocessor in year 2011

ITRS Projections For Year 2011			
Chip Area (mm ²)	800	# gates	1080M
#Signal I/O pads	2000	#Power pads	4000
Memory	90%	Intrinsic gate delay (ps)	2.5
Assumptions			
# Mega cells	20	Placement efficiency	80%

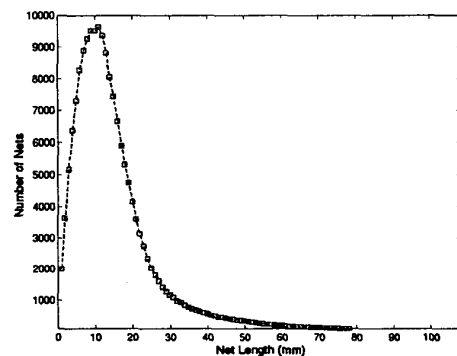


Fig. 1 Global net length distribution for the predefined microprocessor

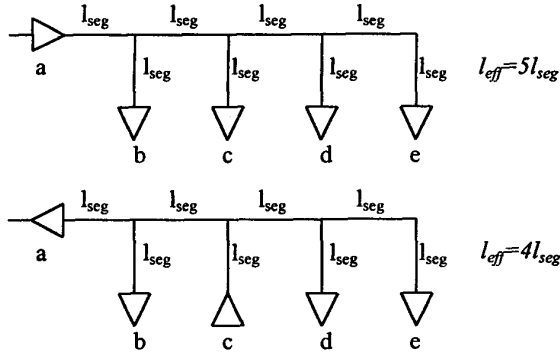


Fig. 2 Wiring net model

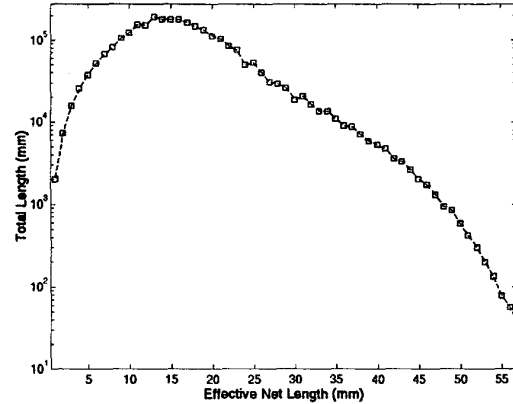


Fig. 3 Effective net-length distribution for the predefined microprocessor

where $f.o.$ is fan out of the net and l_{net} is the net length. Each terminal has equal probability to be the driver. For the case shown in Figure 2, fan out is four and if a, b or e terminals are the driver then the effective length of the net will be $5l_{seg}$ and if c or d terminals are the driver the effective length will be $4l_{seg}$. In another words 60% of all nets with fan out equal to four have effective length of $5l_{seg}$ and 40% of them have effective length of $4l_{seg}$. The same method can be applied to different fan-outs and in this way, total net length versus effective net length can be obtained as is shown in Fig. 3. It is clear that the maximum effective net length is equal to double the chip edge dimension.

Estimating Required Off-Chip Area

Now having effective global net-length distribution, the off-chip area required to route all on-chip interconnects longer than any maximum length can be found:

$$A_{off} = \frac{L(l_{eff} > l_{max}) \times P_{off}}{e_w}, \tag{2}$$

where $L(l_{eff} > l_{max})$ is the total length of all nets having effective length longer than a length (l_{max}) and P_{off} is the off-chip wire pitch and e_w is the wiring efficiency, which is assumed to be 0.5. ITRS has projected wire pitch of $72\mu m$ for printed wiring boards in year 2011 [2].

In order to route on-chip wires through the PWB some I/O pads are required. Figure 4 shows the required number of pads versus maximum on-chip point-to-point interconnect length in two cases. In the first case, all parts of the long nets are placed on board and therefore, each net needs $(f.o. + 1)$ pads. In the second case, only long parts of the nets are routed through PWB and the other parts remain on-chip. The second case requires much fewer pads and also less off-chip area. Figure 5 shows two cases more clearly.

Calculating Total PWB Layers

After finding the off-chip area required for on-chip interconnects, it is necessary to find the number of layers, which should be added to PWB due to on-chip interconnects. The analysis also has to consider the addition of some I/Os to the package by on-chip interconnects routed by PWB.

An algorithm to find total number of PWB layers to route all I/O pads within the footprint of the package on the PWB has been developed [4]. Ground and power I/Os are just connected to the ground and power planes, while signal I/Os should be routed to the outside area of the chip. In order to have smaller via blockage for on-chip wires,

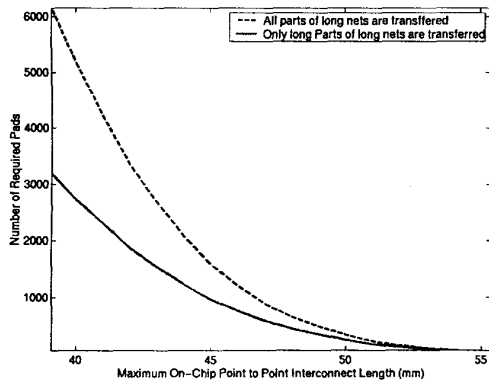


Fig. 4 Required number of pads for on-chip interconnects in the predefined microprocessor

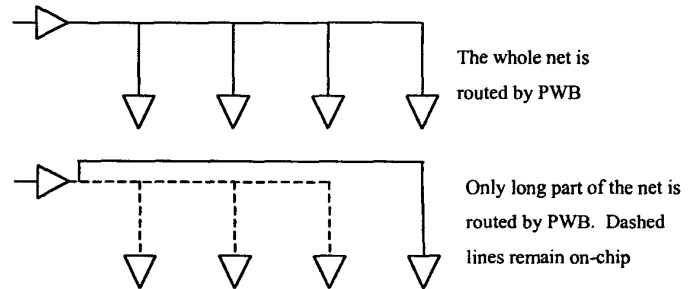


Fig. 5 An alternate method to use fewer pads

upper PWB layers are assigned for routing signal I/Os and lower PWB layers are used to route on-chip interconnects. Assuming that all pads are homogeneously distributed beneath the chip area a simple closed-form relationship between the total number of signal I/Os and number of PWB layers required to route them is found.

Pad pitch, which is equal to via pitch can be found by:

$$P_p = \frac{\sqrt{A_{chip}}}{\sqrt{N_p}}, \quad (3)$$

where N_p is the total number of pads and A_{chip} is the chip area. Number of wires that can be passed through each two adjacent vias defined as lanes per channel is equal to N_l (Fig. 6):

$$N_l = \text{int} \left[\frac{P_p - 2r_v - s}{t + s} \right], \quad (4)$$

where r_v is the radius of the via cross section, t is the wire thickness and s is the minimum spacing between two wires or a wire and a via.

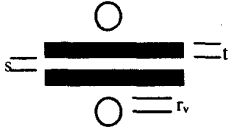


Fig. 6 A channel with two lanes

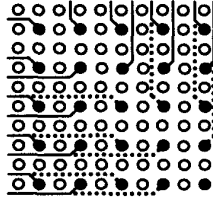


Fig. 7 PWB layers to route I/O pads. Dashed lines should be wired on the next layers

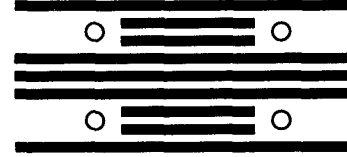


Fig. 8 Via Blockage

Figure 7 shows a quarter of the pad array. Open circles are either power pads or used for on chip interconnects and therefore are not routed to the outside. Signal I/O pads shown by black circles are routed to the closest edge. Since three other quarters are routed the same, they are not shown in Fig. 7. The number of PWB layers required to route signal I/O pads to the outside can be found by

$$N_{pwb} = \frac{\frac{\sqrt{N_{sigpad}}}{2}}{\frac{\sqrt{N_p}}{\sqrt{N_{sigpad}}} N_l} = \frac{N_{sigpad}}{2N_l \sqrt{N_p}}, \quad (5)$$

where $\frac{\sqrt{N_{sigpad}}}{2}$ is the number of signal pad columns from center of the chip to an edge (5 col. in Fig. 7) and $\frac{\sqrt{N_p}}{\sqrt{N_{sigpad}}} N_l$ represents the number of columns that can be routed in each PWB layer (2 col. in Fig. 7, where $N_l=1$).

ITRS projects 2000 signal I/O pads and 4000 ground and power I/O pads for year 2011. It also predicts $36\mu\text{m}$ for t and s in that year although some manufacturers have already reached to this resolution [5]. Via cross section radius is often twice the wire width. The crossed line in Fig. 9 shows number of PWB layers required to route 2000 signal I/O pads for different maximum on-chip interconnect lengths.

Increasing number of pads has also an impact on effective area of PWB layers used for on-chip interconnects. Each pad needs one via to be connected to one of the PWB layers. This via blocks two wires on each layer that it passes through them as shown in Fig. 8. Again the cross-section of each via passing through a layer is assumed to be twice the wire width. The blocked wires can not be used when the distance between vias is small. Therefore each row of passing vias wastes space of two wires. So effective area of each layer is

$$A_{eff} = \left(1 - \frac{2\sqrt{N_v}}{\sqrt{A_{chip}} / 2t}\right) A_{chip} = \left(1 - \frac{4t\sqrt{N_v}}{\sqrt{A_{chip}}}\right) A_{chip}, \quad (6)$$

where N_v is the number of passing vias and t is the wiring width, which is equal to the spacing. $\sqrt{N_v}$ represents number of via rows and $\sqrt{A_{chip}} / 2t$ is the total number of wires in x or y directions on each layer. The above equation shows if 2000 vias pass a PWB layer with wire width of $36\mu\text{m}$ and chip area is 800mm^2 , about 22% area of that layer is wasted. Having the required off-chip area and knowing the effective area of each PWB layer the number of PWB layers to route on-chip interconnects can be found. The dashed line in Figure 9 shows PWB layers required to route on-chip interconnects considering via blockage.

Adding dashed and crossed lines the total number of PWB layers is shown by the solid line in Figure 9. In the case of not using PWB layers for on-chip interconnects about 4 PWB layers are required ($l_{max}=56\text{mm}$). Decreasing the maximum on-chip interconnect length by using board wires, requires more PWB layers. For instance in case of $l_{max}=40\text{mm}$, about 9 layers are necessary (5 layers more) and in order to have $l_{max}=35\text{mm}$, 15 PWB layers are required.

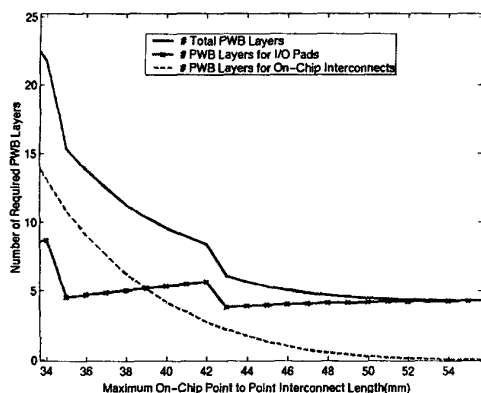


Fig. 9 Number of required PWB layers to reduce maximum on-chip interconnect length for the predefined microprocessor

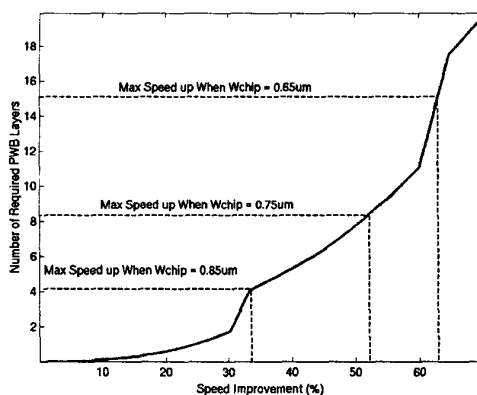


Fig. 10 Number of required PWB layers to increase global clock frequency in the predefined microprocessor

Performance Improvement Using Off-Chip Wires

It is shown that by using a few PWB layers, it is possible to decrease the maximum on-chip interconnect length significantly. Delay for on-chip wires is proportional either to square of the interconnect length or length, depending on whether repeaters are used or not [6]:

$$50\% \text{ delay} = 0.4 R_{int} C_{int} = 0.4 r_{int} c_{int} l^2 \quad \text{No repeater is used} \quad (7)$$

$$50\% \text{ delay} = 2.5 \sqrt{R_{int} C_{int} R_0 C_0} = 2.5 l \sqrt{r_{int} c_{int} R_0 C_0} \quad \text{Optimal number of repeaters is used} \quad (8)$$

where R_{int} and C_{int} are total wire resistance and capacitance and r_{int} and c_{int} are wire resistance and capacitance per unit length. R_0 and C_0 are the output resistance and input capacitance of a minimum size repeater. Since it is often necessary to use repeaters the second case of on-chip interconnects is investigated here.

Assuming that the maximum delay determines the global clock frequency, decreasing the maximum delay results in higher global clock frequency. Figure 10 shows the required number of board layers versus speed improvement. It is interesting that using only two PWB layers the global clock frequency is increased by about 30%, which means 3GHz projected global clock frequency can be increased up to 3.9GHz. There is a limit for speed improvement and it is determined by the time of flight delay of corner to corner interconnects. When all on-chip interconnects slower than those longest interconnects are transferred to the board, moving more interconnects to the board can not improve the performance and theoretically clock frequency has reached to the maximum possible value. Maximum possible speed improvement has been shown in Figure 10 for different on-chip wire widths. In this graph it has been assumed that on-chip and off-chip dielectric constants are equal, and because of recent developments in printed circuit boards it is a reasonable assumption [7]. Both on-chip and off-chip wires are assumed to be copper.

Conclusion

Effective net-length distribution for a projected high-performance microprocessor in year 2011 is found. It shows that there are few nets, which have a receiver very far from the driver. Therefore, moving long parts of those nets to the PWB decreases the maximum point-to-point on-chip interconnect length significantly. Then a model has been developed to find total increase in the number of PWB layers due to transferring some long on-chip interconnects to the board for the same microprocessor. It shows that when no on-chip interconnect is routed by PWB, 4 PWB layers are required to route signal I/Os. When all interconnects longer than 40mm are routed by using off-chip wires, number of PWB layers required to route signal I/Os is increased to 5 while 4 other PWB layers are required for on-chip interconnects. This means that totally 5 layers should be added to PWB in order to decrease maximum on-chip interconnect length to 40mm. Also, the required number of PWB layers for speed improvement is found. It is shown that by adding only two PWB layers the speed can be improved by 30%, which means that the projected global clock frequency of 3GHz is increased to 3.9GHz. Moreover, the maximum possible clock frequency of 4GHz, which is limited by time of flight delay can be achieved by adding 2 more PWB layers, assuming on-chip wire-width of 0.85 μ m.

References

- [1] C. S. Patel, et al., "Low Cost High Density Compliant Wafer Level Package," *International Conference on High-Density Interconnect and systems Packaging*, pp. 262-268, April 2000.
- [2] International Technology Roadmap for Semiconductors, 1999 edition.
- [3] P. Zarkesh-Ha and J. D. Meindl, "Stochastic Net Length Distribution for Global Interconnects in a Heterogeneous System-on-a-chip," *1998 Symposium on VLSI Technology Digest for Technical Papers*, pp. 44-45, June 1998.
- [4] C. S. Patel, et al., "Optimal Printed Wiring Board Design for High I/O Density Chip Size Packages," *Circuit World, Journal of the Institute of Circuit Technology*, pp. 25-27, August 1999.
- [5] C. S. Patel, et al., "An Analysis of the Gap Between PWB Technology and Chip I/O Interconnect Technology, and a new Wafer-Level Batch Packaging Concept," *32nd International Symposium on Microelectronics*, pp. 611-618, October 1999.
- [6] H. B. Backoglu, *Circuit, Interconnections, and Packaging for VLSI*, Reading, MA: Addison_Wesely, 1990.
- [7] R. R. Tummala, *Microelectronics Packaging Handbook*, 2nd edition, Chapman & Hall, 1997.