

Interconnecting Device Opportunities for Gigascale Integration (GSI)

James D. Meindl, Raguraman Venkatesan, Jeffrey A. Davis, James Joyner, Azad Naeemi, Payman Zarkesh-Ha, Muhannad Bakir, Tony Mulé, Paul A. Kohl, Kevin P. Martin
Microelectronics Research Center, 791 Atlantic Drive NW

Georgia Institute of Technology, Atlanta, GA 30332 Phone/Fax: (404)894-9912/0462. E-mail: james.meincl@mirc.gatech.edu

Abstract

In recent years interconnecting devices have become primary limits on the performance, energy dissipation, signal integrity, and productivity of gigascale integration (GSI). Opportunities to address the *interconnect problem* include new materials and processes, reverse scaling, novel microarchitectures, three-dimensional integration, input/output interconnect enhancements, RF wireless interconnects and microphotonics.

Introduction: The Interconnect Problem

The quintessential purpose of an interconnect is *communication between distant points with small latency*[1,2]. A lucid illustration of this key purpose is a graphic whose vertical axis is reciprocal interconnect length squared $(1/L)^2$ and whose horizontal axis is interconnect latency (τ) [2]. Using logarithmic scales on both axes, a diagonal line is a locus of constant $(1/L)^2\tau = r_{int}c_{int}$ or distributed resistance-capacitance product, the principal figure of merit of the large majority of interconnects used for GSI. As illustrated in Figure 1, reducing the distributed resistance-capacitance product moves the diagonal locus toward the lower left corner of the display therefore providing smaller latency for a given interconnect length. However, during the past four decades interconnect scaling has continuously *increased* the distributed resistance-capacitance product thus moving toward the upper right corner of the display and therefore imposing larger latency for a given interconnect length. In stark contrast as shown in Figure 2, scaling of transistors reduces the power-delay product, Pt_d , or switching energy, $E = Pt_d$, of a binary transition therefore moving toward the lower left corner of the power-delay plane to reduce simultaneously both switching energy and delay. This simple dichotomy is the root of the interconnect problem.

Multilevel interconnect networks impose primary limits on the performance, energy dissipation, signal integrity and productivity of GSI. In order to quantify the exploding disparity between the latency of interconnects and transistors consider the comparisons illustrated in Figure 3[3]. The relevant observation is that as semiconductor technology is advancing from the 1.0 μm to the 100 nm generation, the RC delay of a “benchmark” 1.0 mm long interconnect is devolving from 20 times faster to six times slower than transistor intrinsic switching delay. Furthermore, the 1999 ITRS projection for 35 nm technology in 2014 suggests a 2.5 ps transistor delay and a 250 ps RC latency for a 1.0 mm long interconnect [3]. Beyond latency, interconnects present an energy dissipation problem also illustrated in Figure 1. For the 35 nm generation, the binary switching energy of a 1.0 mm long interconnect is 30 times greater than that of a minimum geometry MOSFET. Thus the energy consumption and heat removal problems of GSI are due to interconnects. Finally, the targets for clock frequency, supply current and voltage, number of wiring levels, maximum total interconnect length, and number of bonding pads or input/output interconnects per chip cited in Figure 3 add enormously to expectations for future interconnect capabilities.

Reverse Scaling

An approximate expression for the latency (τ) of an isolated “RC” limited interconnect is given by $\tau = [\rho\epsilon][1/HT][L^2]$. The three factors of this expression represent opportunities to reduce latency through new materials and processes that reduce metal resistivity (ρ) and insulator permittivity (ϵ) ; through reverse scaling of metal height (H) and spacing as well as insulator thickness (T) ; and through novel architectures featuring shared packet-switching interconnect networks that reduce wire length (L) , respectively. Introducing copper conductors and low relative dielectric constant insulators serves to reduce the $[\rho\epsilon]$ factor. In addition, reverse scaling offers relatively timely, low cost, low risk, and high benefit opportunities. The key to reverse scaling is the capacity to predict the complete stochastic wiring distribution $f(L)$ of a macrocell as a function of wire length (L) and number of gates in the macrocell (N) as well as Rent’s coefficient (k) and exponent (p) [4,5] (Figure 4).

In Figure 5, two alternative wiring network architectures are compared. The first architecture shown on the left is restricted to two and only two different cross-sectional dimensions (or two tiers) for eight levels of wiring. It requires two levels of 100 nm and six levels of 540 nm wiring as well as a macrocell area $A_c = 2.34 \text{ cm}^2$ to interconnect the macrocell. The second architecture shown on the right is optimized to use three tiers of wiring in order to minimize cell area. It therefore requires four levels of 100 nm wiring, two levels of 150 nm wiring, and two levels of 300 nm wiring as well as a macrocell area $A_c = 0.70 \text{ cm}^2$. The decisive macrocell area advantage of the three tier architecture is achieved using a methodology whose central feature is demand prediction based upon the complete stochastic wiring distribution defined in Figure 4[4,5].

A second and currently more realistic example of an optimal multilevel network architecture is illustrated in Figure 6. In this case the macrocell consists of an 11.3 million gate random logic network implemented with 100 nm technology using eight levels of copper wiring and operating at a clock frequency of 1.56GHz. If the pitch is chosen a priori to double for every pair of levels, the resulting architecture consists of two levels each of 100, 200, 400, and 800 nm wiring, which require a 1.45 cm^2 area. In contrast, using the previously cited “stochastic wiring distribution” methodology, the optimal wire pair dimensions are 100, 130, 300, and 580 nm, which yield a macrocell area of 0.98 cm^2 or approximately a 32% reduction. Moreover, if 1.5×10^6 optimal repeaters are used macrocell clock frequency can be increased to $f_c = 2.0 \text{ GHz}$ and area reduced to 0.48 cm^2 .

System-on-a-Chip Global Interconnect Architecture

Introducing a heterogeneous version of Rent’s rule (Figure 7), the stochastic global signal net length distribution or number of nets $\{N_{Net}(m)\}$ with m terminals versus average net length $\{L_{av}(m)\}$ is derived. Using this distribution to calculate the total global net length requirement and combining it with models for power and ground wiring area gives the global wiring area (or resource) requirement as shown in Figure 8 [6]. In addition, the most stringent global wiring bandwidth (f_c) requirement imposed by the

H-tree clock distribution network and the crosstalk noise limit are shown in Figure 8 in terms of the physical parameters of the two-level global wiring network [6]. Figure 9 illustrates the allowable design space of an integrated architecture for global signal, clock and power/ground wiring of a representative six million gate heterogeneous system-on-a-chip[6].

The methodology represented in Figures 7-9 enables early projections of key physical parameters of a global interconnect network that *simultaneously* satisfies the primary requirements of a SoC for signal, power, and clock distribution. The compact physical models that serve to implement the methodology offer a convenient opportunity to establish a quantitative guide to detailed design of a SoC. Therefore, the methodology may serve as a useful precursor to final design. Enhancements of this methodology that include, for example, the effects of clock skew and jitter, non-ideal return paths, and simultaneous switching noise are needed.

Three-dimensional Integration

Two prominent approaches to three-dimensional (3D) integration are: 1) a *bonded wafer* approach in which prefabricated wafers are bonded in a stack and provided with area array vertical interconnects[7]; and 2) a *deposited* approach in which thin film transistors are fabricated on upper levels of insulators in a multilevel interconnect network[7]. A principal generic benefit of 3D integration is a reduction in length of the longest global interconnects of a stochastic wiring distribution by $1/S^{1/2}$ where S is the number of strata [8] and a corresponding increase in global clock frequency of $S^{3/2}$ [9]. Stacking chips fabricated with different materials and processes offers opportunities for performance advantages, for example in imaging arrays, mixed signal systems, and displays. Heat removal and input/output interconnect demands of 3D systems appear to be quite challenging.

Input/Output Interconnect Enhancements

Sea of Leads (SoL) technology[10] offers promise of much larger chip I/O bandwidth, time-of-flight global clock frequency with reduced skew, improved suppression of simultaneous switching noise, and enhanced isolation in mixed signal systems as well as reduced cost of packaging, testing, and burn-in via batch processing (Figure 10). This technology utilizes wafer-level-batch-processing of compliant polymer packages, ultra high density ($>10^4/\text{cm}^2$) x-y-z flexible metal leads, and solder-like bumps attached to the lead tips. A short sequence of full wafer SoL batch fabrication processes constituting a "tail-end-of-the-line" (TEOL) are envisaged to follow conventional back-end-of-the-line (BEOL) wafer processing. The further intent of SoL technology is to complete all final electrical testing and burn-in operations prior to wafer dicing that yields *known good packaged die* ready for immediate shipment to customers. The flexible leads are designed to provide sufficient x-y-z compliance to accommodate typical differences in the thermal coefficients of expansion between a silicon chip and the substrate to which it is attached. The need for epoxy underfill is thereby precluded and the possibility of convenient detachment of a chip from a substrate module is enabled.

RF Wireless Interconnects

A key objective of RF wireless interconnect research is to determine the feasibility of *scaling* a code-division-multiple-access (CDMA) or a frequency-division-multiple-access (FDMA) cellular telephone network to the size of a single silicon chip or multi-chip module

[11]. Near-field capacitive RF couplers are proposed to interconnect an array of transceivers through a network of ultra-broad bandwidth (≤ 100 GHz) co-planer waveguides or microwave strip lines for both on-chip and interchip communication.

Microphotonics

Microphotonics aims to develop both on-chip and input/output photonic interconnects[12]. Both bonded and heteroepitaxial III-V compound semiconductor photon emitters and detectors, polycrystalline silicon waveguides and *racetrack* tuned filters that enable wavelength division multiplexing (WDM) for extremely large bandwidth, and polymer waveguides[13] are promising.

Conclusion

A hierarchy of fundamental, material, device, circuit, and system level opportunities to address *the interconnect problem* offers promise of continuing to support the historic exponential rate of advance of semiconductor technology[1,2].

References

- [1] J.D. Meindl, "Scanning the Issue," Proc. IEEE Special Issue on Limits of Semiconductor Technology, vol. 89, No. 3, pp. 223-225, March 2001.
- [2] J.D. Meindl, "Low power microelectronics: Retrospect and prospect," Proc. IEEE, vol.83, pp. 619-635, Apr. 1995.
- [3] International Technology Roadmap for Semiconductors (ITRS), 1999 Edition, SIA.
- [4] J. A. Davis, V. K. De, and J. D. Meindl, "A Stochastic Wire-Length Distribution for Gigascale Integration (GSI) – Part I: Derivation and Validation," IEEE Transactions On Electron Devices, vol. 45, no. 3, pp.580-9, March 1998.
- [5] J. A. Davis, V. K. De, and J. D. Meindl, "A Stochastic Wire-Length Distribution for Gigascale Integration (GSI) – Part II: Applications to Clock Frequency, Power Dissipation, and Chip Size Estimation," IEEE Transactions On Electron Devices, vol. 45, no. 3, pp.590-7, March 1998.
- [6] P. Zarkesh-Ha, J. A. Davis, and J. D. Meindl, "Prediction of Net-Length Distribution for Global Interconnects in a Heterogeneous System-on-a-Chip," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, pp. 649-659, December 2000.
- [7] J.A. Davis et al, "Interconnect Limits on Gigascale Integration (GSI) in the 21st Century," Proc. IEEE, Vol. 89, No. 3, pp.305-324, March 2001.
- [8] J. Joyner et al, "A three-dimensional stochastic wire length distribution for variable separation of strata," IEEE IITC, pp. 132-134, San Francisco, June 2000.
- [9] J. Joyner, et al, "A global interconnect design window for a three-dimensional system-on-a-chip," Proceedings IEEE IITC, pp. 154-156, June 2001.
- [10] A. Naeemi et al, "Sea of Leads: a disruptive paradigm for a system-on-a-chip (SoC)," IEEE ISSCC, pp. 280-281, San Francisco, Feb. 2001.
- [11] M.F. Chang et al, "Multi-I/O and reconfigurable RF/wireless interconnect based on near field capacitive coupling and multiple access techniques," Proc. IEEE IITC, pp. 21-22, June 5-7, 2000.
- [12] D.A.B. Miller, "Rationale and Challenges for Optical Interconnects to Electronic Chips," Proc. IEEE, Vol. 88, No. 6, pp1 728-749, June 2000.
- [13] A.V. Mule, S.M. Schultz, E.N. Glystis, T.K. Gaylord, J.D. Meindl, "Input Coupling and Guided-wave Distribution Schemes for Board-level Intra-chip Guided-Wave Optical Clock Distribution Network using Volume Grating Coupler Technology", IITC, Session 6, San Francisco, CA, June 4-6, 2001.

We gratefully acknowledge DARPA, contract F33615-97-C-1132, the SRC, contract 448:048, and the Interconnect Focus Center (IFC) funded in part by MARCO contract B12-M00 for their generous support.

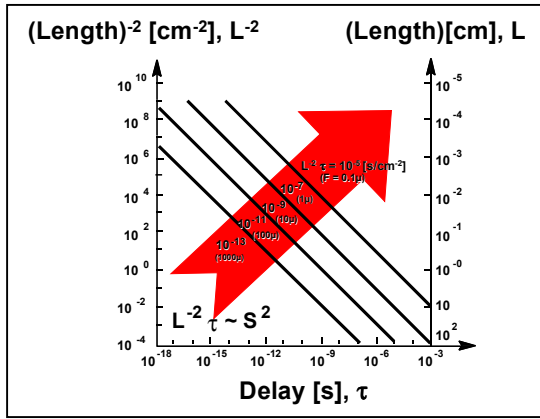


Fig. 1. Reciprocal interconnect length squared $(1/L)^2$ vs latency (τ). Diagonal lines are loci of constant $(1/L)^2\tau$ product or distributed resistance-capacitance ($r_{int}C_{int}$) product.

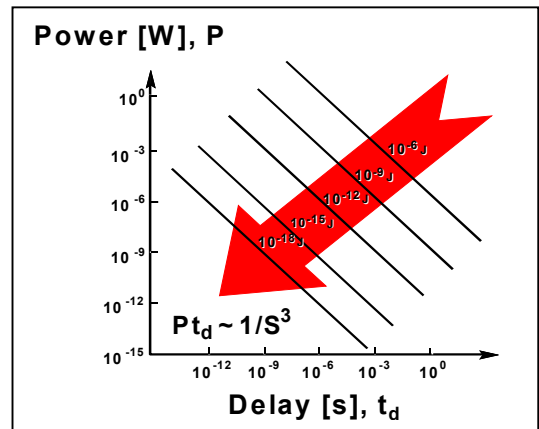


Fig. 2. Average signal power transfer (P) vs switching delay (t_d) for a MOSFET binary transition. Diagonal lines are loci of constant power-delay product or switching energy $E = Pt_d$.

Interconnect Performance Requirements

| | Technology Generation | | |
|---|-----------------------|-------------|----------------|
| | 1.0 μm | 100 nm | 35 nm |
| MOSFET Switching Delay | ~ 20 ps | ~ 5 ps | ~ 2.5 ps |
| Interconnect "RC" Response Time ($L_{int} = 1 \text{ mm}$) | ~ 1 ps | ~ 30ps | ~ 250 ps |
| MOSFET Switching Energy | ~ 300 fJ | ~ 2 fJ | ~ 0.1 fJ |
| Interconnect Switching Energy ($L_{int} = 1 \text{ mm}$) | ~ 400 fJ | ~ 10 fJ | ~ 3 fJ |
| Clock Frequency | ~ 30 MHz | ~ 2-3.5 GHz | ~ 3.6-13.5 GHz |
| Supply Current ($V_{dd} = 5.0, 1.0, 0.5 \text{ V}$) | ~ 2.5 A | ~ 150 A | ~ 360 A |
| Maximum Number of Wiring Levels | 3 | 8-9 | 10 |
| Maximum Total Wire Length per Chip | ~ 100 m | ~ 5000 m | () |
| Chip Pad Count | ~200 | ~ 3000-4000 | 4000-4400 |

Fig 3. ITRS projections for switching delay, switching energy, clock frequency, total chip current drain, maximum number of wiring levels, maximum total wire length per chip and chip pad count for 1.0 μm , 100 nm, and 35 nm technology generation.

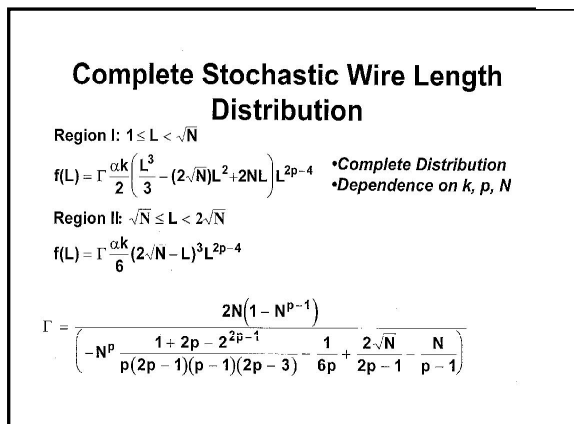


Fig 4. Complete stochastic interconnect length (L) distribution $f(L)$ for a random logic network of N gates where k and p are Rent's coefficient and exponent, respectively. The first equation applies to shorter wires and the second to longer wires in the distribution.

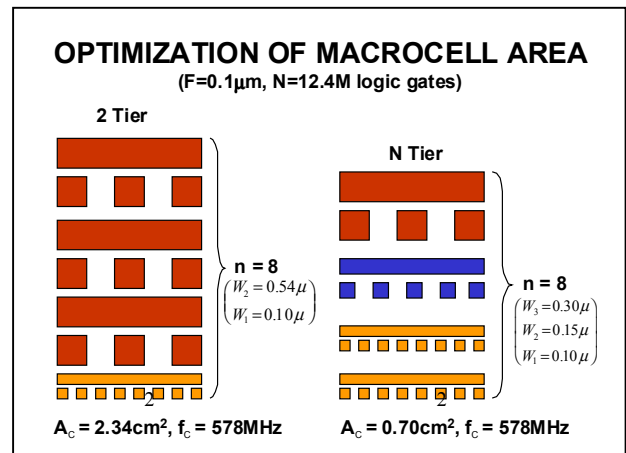


Fig 5. Comparison of wiring limited macrocell areas required for a two tier versus an N tier (for optimal $N = 3$) multilevel interconnect network architecture.

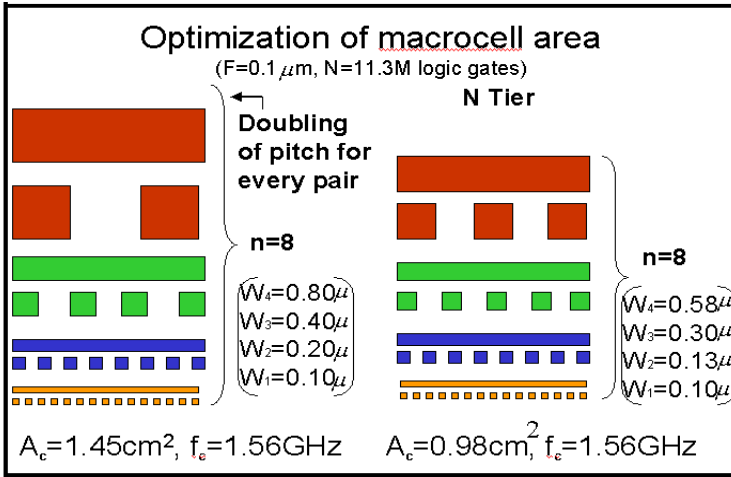


Fig. 6. Comparison of wiring limited macrocell areas required for a two tier versus an N tier (for optimal N = 3) multilevel interconnect network architecture.

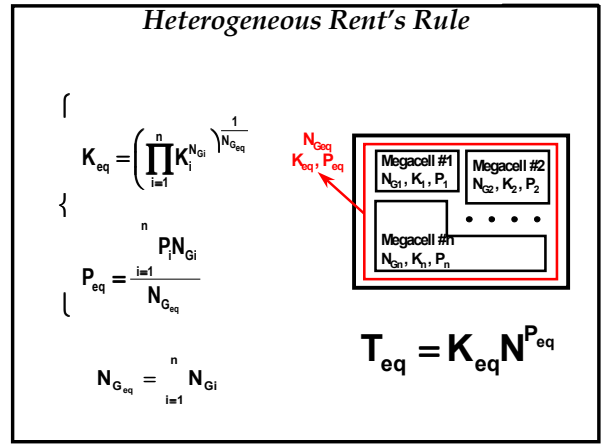


Fig. 7. Definition of heterogeneous Rent's rule that applies to a heterogeneous set of megacells #1 through #n [12] comprising a system-on-a-chip.

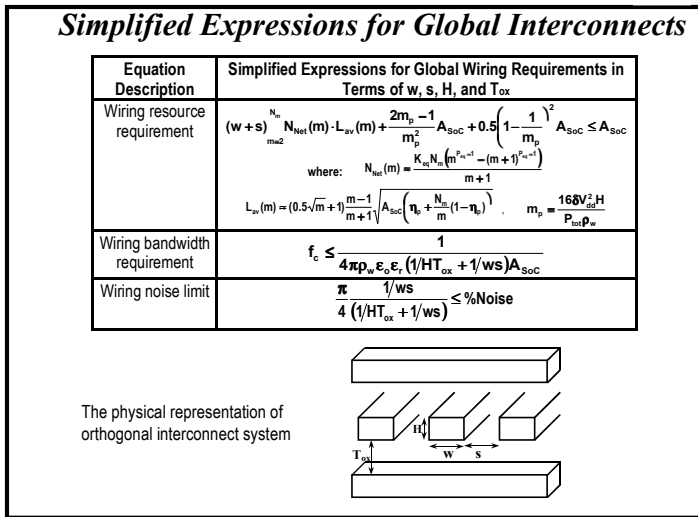


Fig. 8. Summary of complete set of requirements to be imposed on global signal, power, and clock distribution networks expressed in terms of the geometry of the two global wiring levels

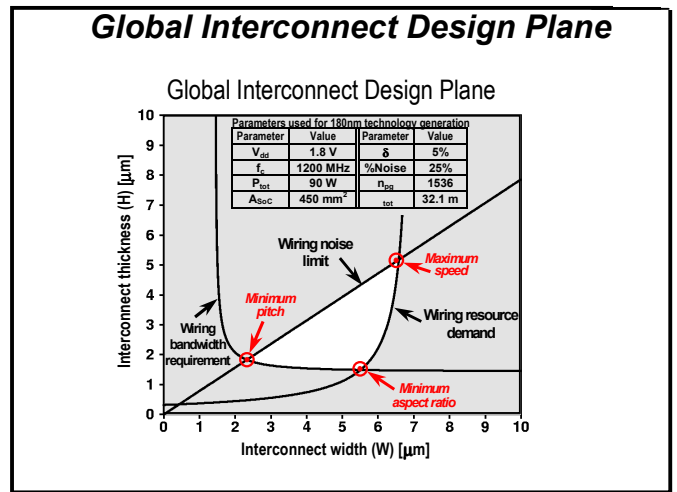


Fig. 9. Global interconnect design plane plotting interconnect thickness H versus width W for the interconnect requirements summarized in Figure 8 as applied to a SoC consisting of 20 heterogeneous megacells including a total of approximately six million transistors

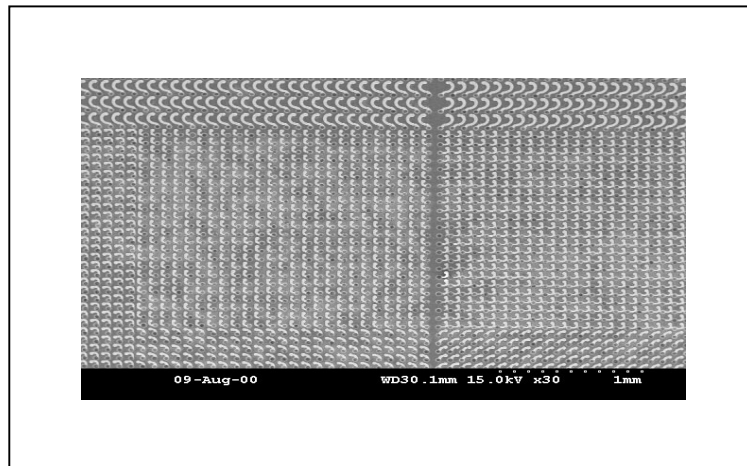


Fig 10. SEM of a Sea of Leads with a density of 12,000 per cm²