The Effect of Zero Resistance Interconnect in Silicon ICs

Jeff Yetter, Yetter Consultancy   Jan. 12, 2009

Overview

Modern Si ICs pay a penalty for interconnect resistance in operating frequency, power dissipation, power delivery, design complexity and ultimately in manufacturing cost in the form of extra layers for power and signal interconnect to compensate for IR and RC effects of interconnect. The purpose of this paper is to take a qualitative look at a hypothetical Si IC with interconnect that is free of resistance. Such a technology would present designers with a new toolkit for the design of vastly improved ICs and a challenge to overcome new problems that would otherwise reduce the inherent benefit of the zero resistance technology. Such problems however, in this author's opinion would be "luxury problems" with solutions that are far simpler than those posed by traditional Cu and Al interconnect.

Forward

In the following pages I will look at five common limiters to the continued scaling evolution of Si ICs, all of which stem from the limitations of resistive interconnect. These five are: 1) Power delivery, 2) Clock distribution, 3) Global signal interconnect, 4) Power dissipation and 5) Local interconnect. This will provide a basis for further quantitative analysis of these effects. Finally, I will propose what new challenges will need to be resolved to take full advantage of zero resistance interconnect and discuss means to resolve those challenges.

Power Delivery

Modern (45nm to 60nm) ICs contain nine or more layers of metalization for distribution of power, clocks and signals. In most cases, two of these layers are allocated primarily for power alone, and a significant portion of the remaining seven are allocated for power distribution. In addition, a portion of the silicon substrate, approximately 10 – 15% is allocated for the inclusion of local bypass capacitance. The additional interconnect layers add to manufacturing cost in the form of extra masks, lithography steps and yield reduction. In addition, power distribution has become a major design discipline in its own right, affecting time to market, circuit density, signal integrity, parasitic capacitance (and therefore speed) and overall design cost to create and verify the power grid. The power grid itself can account for a significant portion of the total chip power consumption in the form of DC IR losses, CV^2f (dynamic) power attributable to voltage ripple, leakage in the bypass capacitor structures and reduced effective (worst case) voltage applied to active structures resulting in over design inefficiency. The effect of self- and mutual-inductance, while not negligible in traditional resistive power distribution, are by far secondary to the IR and RC effects.

Absent resistance, IR, RC and RL effects of power distribution would disappear. Only the source impedance of the regulator and drain and source resistance of the active devices would remain. These however are very localized effects and account for only a small portion of the total loss in a traditional interconnect system. The new landscape would however become an undamped or underdamped LC world. The implications of this will be discussed further at the end of this paper but suffice it to say that electronics offers us an ample solution space for any related complications. With that in mind we can expect the following:

- Reduced levels of interconnect, potentially from 9 to 6 layers for the same amount of signal connectivity.
- Reduced investment in silicon area for bypass cap structures and circuit overdesign to compensate for IR and RC effects of power distribution
- Reduced power (in the 10-20% range) for losses in the interconnect and reduced leakage in bypass caps and overdesigned gates
- Improved development cost and time to market penalties associated with power grid engineering

Some portion of these benefits may have to be reinvested to resolve issues caused by the underdamped distribution, but the gains can be expected to be significantly positive.

Clock Distribution

Most (but not all) modern ICs are still synchronous devices. They depend on a global (chip wide) clock distribution with phenomenal gain. A state-of-the-art clock distribution system starts with a point source, typically a phase locked loop that multiplies an external frequency reference, and a high gain buffer. The source output is then routed through a balanced interconnect tree with the objective that the signal reaches all terminal points simultaneously. The tree usually contains some amount of buffering. The terminal points are then fed into high gain delay locked loop circuits ideally capable of zero-delay amplification, then subsequently routed and rebuffered before being ultimately applied to latches. In an ideal world all latches receive a perfectly simultaneous image of the clock.

In the real world of resistive interconnect simultaneous delivery is unobtainable. Variations of routing delay are caused by linewidth variation, uncontrollable parasitics (capacitive), mismatches in the tree’s buffering components and other effects. Further distortion is introduced in the from of jitter and drift in the delay locked loop circuits and mismatched terminal buffering. The total deviation from the ideal simultaneous delivery is called clock skew. Clock skew has many adverse effects on circuit operation, but the dominant ones include a reduction of clock frequency and the creation of race conditions wherein communications between circuits operating on different clock terminals is disrupted. Circuit designers have devised clever methods to try to minimize the effect of clock skew at a high cost in design productivity, but ultimately the price is paid through a reduction of the clock frequency. Specifically, twice the value of the clock skew (typically at least a few hundred picoseconds in a well-designed clock system) must be subtracted from the minimum clock period obtainable by the same chip design with an ideal clock distribution. It should also be noted that these effects do not benefit from further scaling of technology to smaller dimensions. **This may well be the single biggest obstacle to improvements in clock frequency when scaling past the 90nm –120nm technology, and probably accounts for the reason that we are no longer seeing the vast frequency improvements in microprocessors that we were so accustomed to just a few years ago.**

Zero resistance interconnect has a lot to offer this problem. For large designs, a distribution tree will still be required, but free of RC effects a vastly improved result will be obtained. Skew components from unbalanced parasitics and line width variations will become negligible. Designers will have the opportunity to use minimum dimension linewidths in the distribution, resulting in much lower capacitance and therefore will require far less gain and buffering. This will have the effect of reducing buffer mismatch from the tree. In the best case even the delay locked loops may be eliminated and replaced with simple buffers, or even nothing at all. The jitter component disappears. All these combined, it is a reasonable goal to reduce total clock skew from a few hundred picoseconds to a few 10s of picoseconds.

To put that in perspective, a 2GHz clock has a period of 500pS. 200pS (a reasonable state of the art global skew number) is 40% of the clock period. A 10-fold reduction in clock skew to 20pS should then support a 50pS clock period, or a 20GHz microprocessor.

Clearly limiters other than clock distribution must be resolved to achieve that result, but the elimination of resistance would at the very least move the single top frequency limiter to a position far down the list. We could expect that to pay immediate dividends in the scaling frequency as process and lithography enable ever-smaller features.

All that in mind, we should expect at least the following:

- Vastly reduced global clock skew in synchronous circuits, and a potential attendant increase in operating frequency.
- Improved design productivity by eliminating or simplifying the active balanced tree and the attendant reduction in design cost and time to market.
- A reduction in power from the clock distribution (I will estimate here without justification about a 20% reduction in overall chip power for a large device such as a microprocessor).
- A minor reduction in silicon area and routing resources.

Global interconnect

In IC design parlance Global interconnect is loosely defined as the sum of wires that route between major blocks of a chip. These wires range in length from about 1000μ to 1.5 cm. Delays created in global routes range from a few hundred picoseconds to near a half-nanosecond (entire clock cycles!). This delay is entirely due to parasitic RC. Since about the 250nm generation global signal propagation has gained assistance using repeater stations that repower signals en-rout to their destination. This works out up do a distance of about .5cm, but beyond that far more elaborate solutions are required. Signals in that distance range may require latched, or pipelined repeaters. When that becomes necessary the repeater configuration becomes visible to the microarchitecture, and in a nutshell creates a design complication of huge proportion. This describes the state of the art.

Insertion of repeater stations presents a particular kind of strain on design productivity. This is because a mere schematic cannot represent the function and placement of a repeater, it must be factored into the physical floorplan. Global timing analysis must then comprehend the buffer insertion delay. The buffers themselves require power delivery thus affecting the power grid. When a latching repeater must be used the insertion must be accommodated by a logic designer with potentially large impact to other parts of the design. Global routing of a large fast chip is, simply put, a designer's nightmare.

To state the obvious, routing delays, be they repeated or not must be part of an overall circuit delay budget. A typical rule is that a global signal is allocated a half-clock cycle with at most one gate delay either at the source or at the destination. This appears to the logic as wasted time. Longer routes requiring latching repeaters require allocation of pipeline stages and are best identified early in the design cycle when a viable floorplan does not yet exist. Adding pipeline stages later in the design causes huge rework of logic and considerable reverification and causes seemingly unbounded design delays.

Power consumption for global routes is largely CV^2f, where C is the line capacitance and f is the switching frequency, plus the switching power of the repeater. Thus, power is largely independent of line resistance.

Nonetheless, zero resistance interconnect has a lot to offer to this problem.

Without resistance parasitic RC delay is eliminated. For a 2cm chip, repeaters and latching repeaters should be eliminated. Routing delays should shrink to nearly zero. The savings can be captured a number of useful ways: Clearly a savings in cycle time can be captured. Alternatively, designers can combine logic delays in paths with global routes, which can further reduce cycle time or reduce pipeline stages. Logic designers, floorplaners and power/clock grid engineers no longer need to consider route lengths and repeater insertion which vastly reduces design time and cost and improves time to market.

All that being said, we can expect at least the following:

- A reduction in cycle time of at least 50% for paths that include global routes
- Alternatively, a reduction of one-half to one pipeline stage for paths that include global routes.
- A vast improvement in design productivity and time to market for large fast designs.
- A modest reduction in power consumption and silicon area due to the elimination of repeaters, and reduced capacitance resulting from the use of finer interconnect (closer to the minimum design rule).
- Improved frequency scaling as designs are ported to denser processes.

Power Dissipation

There are three dominant components of power dissipation in traditional ICs (in order of importance):

1. CV^2f: This is the power required to charge and discharge capacitive loads on an IC. It is often referred to as dynamic power. C is the effective capacitance of each switched node, V is the effective operating voltage and f is the effective switching rate of each switched node. A node can be a local or a global signal, it can be the gate, source or drain of a MOSFET or it can be a clock interconnect or fanout.
2. Switching power: As a gate's input transitions from one rail to the other it crosses through a point where both NMOS and PMOS FETs conduct, causing a DC current flow through inverters and gates. Gates with slow input transitions contribute most heavily to this component. Large (high output drive) buffers contribute excessively to switching power and
3. DC losses: This is a catch-all category with many components. Some circuit types, such as differential circuits, analog circuits (including PLLs and DLLs) and "psuedo NMOS" circuits present constant DC draw. Other components include FET (D-S) leakage, gate leakage to substrate and non-fatal defect draws.

Now look at each of these in the presence of zero-resistance interconnect:

Dynamic power (CV^2f): This power accounts for roughly 65% of a circuit's draw. Let's consider f as fixed for this discussion (although our goal is clearly to increase f). A linear power reduction is achieved by reducing C. Zero resistance interconnect will serve this purpose well. First of all, most long interconnects will reduce to near minimum width. In contrast, present technology requires wide (2-3X minimum design rule) interconnect to reduce overall resistance and line width variations that further increase worst case resistance. An immediate savings of roughly 50% is obtained in the capacitance of these routes. Consequently, smaller circuits (less gain) are therefore required to drive those interconnects, resulting in lower gate capacitance, also in the range of 50%. Shorter interconnects will benefit also, although at a lower percentage. Very local interconnect will receive negligible benefit in this regard. Clock distribution systems, as discussed earlier should reduce their capacitance by greater than half. This is significant since they can represent up to 35% of the total chip capacitive load.

As a conservative estimate, C will be reduced by about one-half in a resistance free design.

Voltage reduction: It would seem at first that interconnect resistance would have at most a minor effect on circuit operating voltage, but it deserves a closer look. Circuits are designed to maintain their performance worst case voltages. By eliminating IR and RC drops in interconnects, worst cast and nominal voltages become closer together. Since the voltage term is squared in the power equation this will be significant. A good estimate is that supply voltage can be reduced by about 15% with a well designed zero resistance power distribution, resulting in a 28% reduction in power consumption.

Switching Power: Crossover switching in gates and buffers accounts for roughly 25% of the total dissipation. The amount of the draw is proportional to the voltage slew at the gate's input. Slower input signal transitions cause gates to spend more time in the crossover region. For local logic circuits this is typically small to negligible since local signals tend to switch fast. Global signals and clocks however, due to inherent RC delays cause appreciable switching power at their destinations. In the absence of RC dominated slew rates (dV/dt) switching power will be greatly reduced. Without a lot of justification, I would like to propose a 60% reduction goal in switching power for a chip where R=0.

DC losses: These account for roughly 10% of a chip's power draw. It is not likely that we can eliminate DC circuits just because we have zero resistance interconnect. And leaky bypass capacitors will likely still be required, although to a lesser extent. However, all DC effects will be reduced by the square of the voltage reduction (as they are essentially V^2/R in nature). My estimate according to previous criteria is a 15% voltage reduction.

Power summary:

All other things being equal, we have developed a simple power dissipation model as follows:

$65\% * .5 * .85^2$	dynamic power (includes reduction in clock distribution)
$+ 25\% * .4 * .85^2$	switching power
$+ 10\% * .85$	DC losses
= 0.379	roughly 62% of the power has been saved

Local Interconnect

Local interconnect accounts for most of the lower layers of metal on an IC and spans anywhere from a few microns to hundreds of microns. It typically uses minimum or near-minimum design rules and does not require repeaters. Examples include functional units such as ALUs, standard cell circuits such as control logic, word- and bit-lines in memory arrays and register files and cell local interconnect.

It would appear on the surface that zero resistance interconnect would not benefit such connections, but further consideration reveals something interesting: I would argue that these interconnects are not currently frequency limiters at today's level of technology. I cite as evidence the fact that since the 120nm technology generation no appreciable frequency gains have been achieved in microprocessors. (The first 2GHz microprocessors were fabricated at 120nm and this remains close to the standard frequency in the 60nm generation). To be sure, the recent scaling has reduced power consumption and manufacturing cost and has enabled multi-core products but overriding limiters in clock distribution, global interconnect, design complexity and power distribution have prevented the technology from delivering the frequency improvements that were a standard feature of previous generations. Gate delays are not to blame because transistor switching speed scales well with geometry.

Having said all that, there are still many ways that local circuit delays can be addressed by design. One distinct possibility is deeper pipelines resulting in fewer gate delays per cycle. Certainly the dimensional scaling of technology enables wider logic trees and increased redundancy in logic, which also reduces gate delays. While these techniques have been available to designers all along, their application was irrelevant in light of the other more dominant problems that prevent frequency scaling.

In summary, zero resistance interconnect does not in itself bring an appreciable frequency benefit to local circuits. But it will put dimensional scaling back on the track of providing the frequency increase that was so abundant prior to the 120nm generation. Arguably, in the best case, the first application of zero resistance interconnect to the global issues will afford a triple generation leap if applied to a 60nm or 45nm generation.

Speculation on new challenges

The most likely new challenge faced by designers will be to learn to design in an undamped LC world. This is unknown territory in that no previous examples exist from which to draw experience, and no serious analysis of this domain has been performed to my knowledge. The risk of course is that ICs will have the tendency to become oscillators.

Of course, this can be resolved by providing a damping resistance to the power grid however at the cost of mitigating the best benefits of the technology, but it does bound the problem. A better engineering solution may be to tune the power grid to be free of resonance in the clock frequency band. Distributed regulation should also be an effective solution, particularly if the regulators had active shunt damping capability. It is likely that this approach would cost far less area and scarcely more power than the leaky bypass capacitors they replace. This remains an area for further study and an opportunity for engineering contribution.

Another area to watch is slew rate, or di/dt on signal interconnect. The hazard there is mutual inductance, or inductive coupling. This has been largely negligible in the RC world of Cu and Al interconnects. Fortunately, there are effective solutions to this problem. Simply scaling down signal drivers (which are typically over designed to compensate for RC delay) will be effective and will have the added benefit of reducing crossover current. For very high current nets such as the legs of the clock distribution tree, some shielding may be required. The cost of shielding can be offset if it is combined with the power distribution nets. For very high performance applications differential signaling may be employed.

Final note

This paper draws most of its examples from the microprocessor style of IC. However, the arguments put forth should be equally valid for any large digital design, including memories, SoCs (Systems on a Chip), and system components such as backplane routers and graphics chips. We should expect the arguments to be valid as well for smaller designs, however with somewhat less total benefit.