High bandwidth, Low Latency Global Interconnect

Christer Svensson* and Peter Caputa
Dept. of Electrical Engineering, Linköping University, 581 83 Linköping, Sweden.

ABSTRACT

Global interconnects have been identified as a serious limitation to chip scaling, due to their limited bandwidth and large delay. A critical analysis of intrinsic limitations of electrical interconnect indicates that these limitations can be overcome. This basic analysis is presented, together with design constraints. We demonstrate a scheme for this, based on the utilization of upper-level metals as transmission lines. A global communication architecture based on a global mesochronous, local synchronous approach allows very high data-rate per wire and therefore very high bandwidth in buses of limited width. As an example, we demonstrate a global, 250µm wide bus with a capacity of 160Gb/s in a nearly standard 0.18µm process.

Keywords: Interconnect, low latency, high bandwidth, global

1. INTRODUCTION

Integrated circuit interconnect has been considered a showstopper for process scaling because of their RC delays[1-5]. However, it was early noticed that the RC-delay could be considerably reduced by utilizing upper level, thicker metals for long interconnects[3]. It was in fact predicted that global synchronism could be maintained until the feature size reached 0.3µm. Today, we know that high speed interconnects must be described by models which include not only R and C but also inductance and skin effect[4-6]. The first thought is that this will make the situation worse, but we found that it is not so. In this paper we will show that well designed, highly lossy, long interconnects may show reasonable delays of the order of twice the delay compared to the velocity of light delay, and allow high data rates.

Contemporary VLSI developments can be characterized as System on Chips (SoC). A SoC is often designed as a number of predefined blocks, or IP-cores, for example processors, DSP's or memories, which are integrated into one chip. A key issue in SoC design is how to facilitate an effective communication between the various IP-cores. Traditionally, this communication constitutes custom buses and glue logic, often demanding large design and verification effort. A new trend is to utilize Networks on Chip (NoC)[7-10], both in order to mitigate the design and verification effort and to increase the available communication bandwidth between the IP-cores. In the present paper we will use the NoC model for SoC communication to demonstrate an efficient utilization of upper level metal interconnect.

2. BASICS OF WIRES

The performance of wires or bundles of wires (buses) can be described in terms of four parameters, delay (or latency), maximum data-rate, crosstalk and power consumption. We will first discuss wire modeling in general and then return to these four parameters. The most general wire model is the transmission line model (The telegraph equation)[4]. This model is only valid for wires with a well defined return path, as for example twisted pairs, coaxial cables, striplines or coplanar waveguide, see fig. 1. As we are interested in high-performance wires, we will only consider such well-defined lines in the following, mainly microstrip. The wire is thus described by its transfer function, H, and characteristic impedance, Z_c (complex), calculated from inductance, l, capacitance, c, and resistance, r, per unit length, wire length, L, and including skin effect[6,11]. H is the transfer function of a wave moving through the wire, and is given by:

$$H = e^{-\sqrt{(j\omega l + r)j\omega c}L} \quad (1)$$

* chs@isy.liu.se, www.ek.isy.liu.se

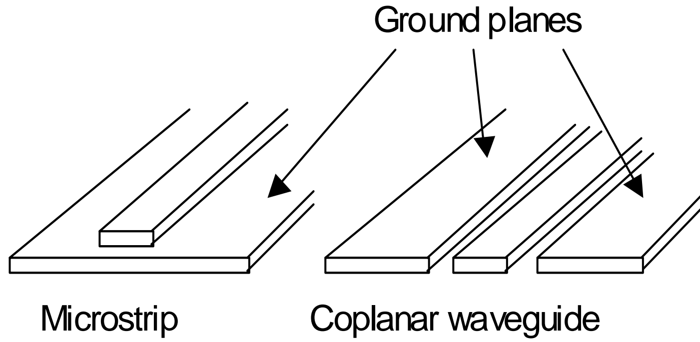

Fig. 1. Microstrip and coplanar waveguides.

Z_c is the relation between voltage and current in each wave on the wire (one wave in each direction):

$$Z_c = \sqrt{\frac{j\omega l + r}{c}} \tag{2}$$

For high frequencies Z_c approaches $\sqrt{l/c}$, which we will term Z_0. Also at high frequencies, skin effect should be included in r. r can then be expressed as:

$$r = r_{DC} + r_s(1+j)\sqrt{\omega} \tag{3}$$

where the second term describes skin effect (the j-term describes the inductive part of the skin effect). Also dielectric loss may be included in these formulas, through a complex and frequency dependent dielectric constant in c[11], but dielectric loss is not considered important inside chips.

If the wire is connected to a driver with output impedance Z_S and loaded by an impedance Z_L, we may express the total transfer function from driver emf to voltage over the load as G:

$$G(\omega) = \frac{2Z_L H}{Z_L\left(1 + H^2 + \frac{Z_s}{Z_c}\left(1 - H^2\right)\right) + Z_c\left(1 - H^2 + \frac{Z_s}{Z_c}\left(1 + H^2\right)\right)} \tag{4}$$

Eq. (4) is obtained from circuit equations, combined with the two waves on the transmission line[4]. We may note, that for perfect terminations, $Z_L=Z_S=Z_C$, G=H/2. Further, for $Z_S=Z_C$ and $Z_L=\infty$, G=H, that is we get the full amplitude from the source appearing at the far end of the wire.

In order to better understand the wire behavior in various circumstances, it is useful to analyze the step response of its transfer function, h, which can be found from the inverse Fourier transform of H[11]. In fig. 2 we show h for the case of 50% wire attenuation. We note that the step response is characterized by an initial delay, which corresponds to the velocity of light delay, a step with limited height, which height corresponds to the wire attenuation, and finally a slow rising path approaching 1, which corresponds to RC charging of the wire. Taking skin effect into account will make the step slower (the frequency-dependence of the skin-effect gives rise to signal distortion). For most integrated circuit wires, the attenuation is very large, making the RC-charging (RC-behavior) dominate the behavior, and they have an open far end, making Z_L purely capacitive (C_L). For this case Eq. (4) is reduced to the Elmore delay expression[12], describing the wire with a π-circuit, consisting of a series resistor with the wire DC-resistance and two capacitors, half the wire capacitance each:

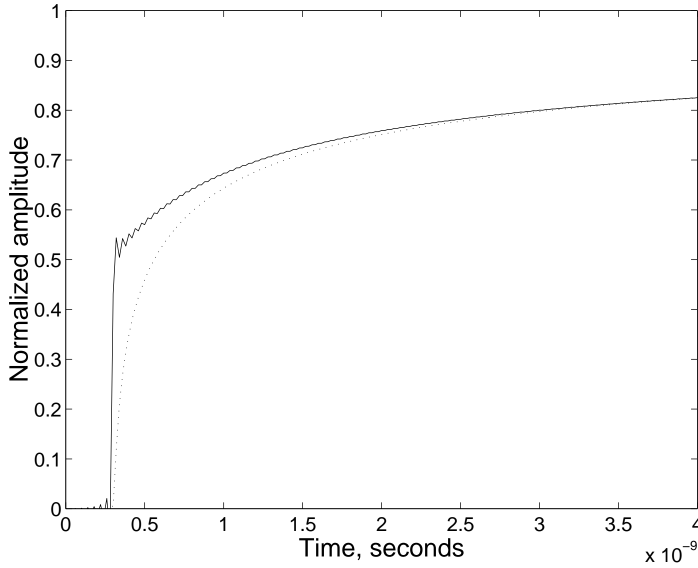

Fig. 2. Step response of a wire transfer function, solid: without skin effect, dashed: with skin effect (The weak ringing is a mathematical artifact).

$$t_{Elmore} = \left(R_S \left(C_S + C_w + C_L \right) + R_w \left(\frac{C_w}{2} + C_L \right) \right) \ln 2 \tag{5}$$

where R_S and C_S is the source resistance and capacitance and R_w and C_w is the wire resistance and capacitance respectively. This is then the traditional view of integrated circuit wires, characterized with large delays (much larger than delays related to the velocity of light). For the opposite case, wires with relatively small attenuation, the step in fig. 2 dominates (LC behavior), and the wire delay can be considered to be:

$$t_d = \frac{L}{v_d} \tag{6}$$

where v_d is the velocity of light in the actual dielectric used, $v_d=v_0/n$, where v_0 is the velocity of light in vacuum and n is the refractive index, given by $n = \sqrt{\varepsilon_r}$, ε_r being the relative dielectric constant. Low loss wires thus have the smallest possible delay allowed by physics. The borderline between the two cases, RC-behavior and LC-behavior occur for an attenuation of 50% (considering that we measure delay from the step launch time to the time at which the signal reaches 50% of its final value at the far end). Using the "classical" loss formula[4]:

$$\left| H \right| \approx e^{-\frac{rL}{2Z_0}} \tag{7}$$

gives the constraint for the wire to behave as a transmission line:

$$\frac{rL}{Z_0} \leq 2 \ln 2 \tag{8}$$

Returning to the full expression for G, eq. (4), we may calculate its step response, g, in the same way as we calculated h. This was done for $Z_S=Z_0$, $Z_L=\infty$ and 50% loss, and shown in fig. 3. We note that g(t) (or the step response S(t)) is in fact much steeper than h(t). Particularly, h(t) show a very long tail which is very detrimental for the data-rate, as it gives rise

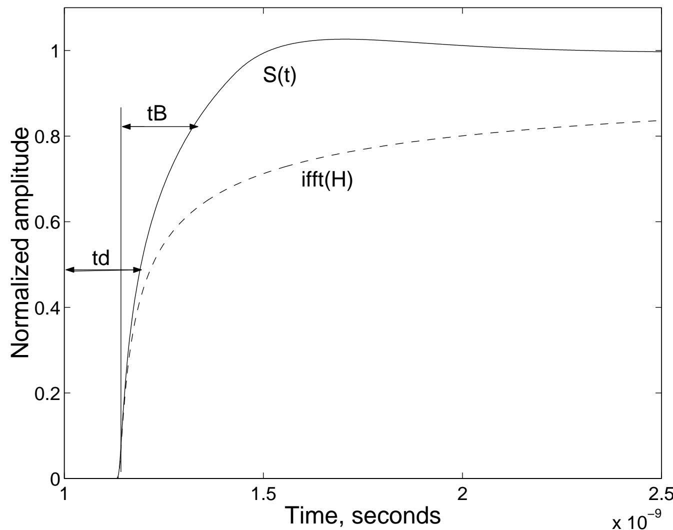

Fig. 3. Theoretical step response, S(t), of an open wire driven by a terminated driver, G, compared to the step response of the wire transfer function, H.

to intersymbol interference (ISI). ISI means that one symbol affects subsequent symbols through the long tail. The fact that g(t) is much better in this respect is caused by the difference between the actual characteristic impedance of the wire, Z_C and its high frequency value, Z_0, causing an enhancement of the input voltage at the input edge. This is an important effect, which we utilize in this work.

Let us now return to the four performance parameters of the wire.

The delay, or latency, of the wire is thus given either by the velocity of light of the actual dielectric used or by the RC delay. Velocity-of-light delay, eq. (6), is valid for wires with loss less than about 50%, i.e. for wires fulfilling the constraint eq. (8) above. On the borderline, we need to add part of the wire risetime to the delay, making it about double L/v_d. For these wires, repeaters do not improve delay, as delay cannot be less then velocity-of-light delay anyway. For wires with larger resistance, the delay is given by the RC-delay formulas, eq. (5), often considerably larger then the velocity-of-light delay. We will only consider wires with velocity-of-light delay in the following.

The maximum data-rate of a wire is controlled by its step response, S(t). In a lossy wire the data-rate is limited by the intersymbol interference (ISI) imposed by the transfer function of the wire. The effect of ISI can be quantified as the minimum eye-opening of the eye-diagram of the received signal. This minimum eye-opening (the smallest difference between a “0” and a “1”) occurs for a single zero and a single one, and is given by 2S(T)-1, where S(T) is the step response of the transfer function and T=1/B is the data period of a binary data stream with bit-rate B[6]. For a lossless transmission line B is infinite, but with loss the value is lower. It turns out that the step response of an open wire, driven by a driver with output impedance Z_0 or less, is better than the step response of the transfer function itself, as discussed above in connection to fig. 3.

As we are interested in the time at which S(t) is relatively large (typical 0.8, t_B in fig. 3), the two curves in fig. 3 yields very different data-rates. We may interpret the upper curve as a result of pre-emphasis, i.e. the input voltage to the wire is enhanced in the beginning of the bit[13] (in this case occurring as a result of a fixed driver impedance compared to a frequency-dependent characteristic impedance of the wire). This effect can be further emphasized by choosing a lower driver impedance, see the example below.

If we consider a bundle of wires, we also need to consider crosstalk. For a simple bundle of parallel wires, the worst case crosstalk occurs when the two neighboring wires have data edges opposite to the edge on the actual wire. The amount of crosstalk is controlled by the mutual inductances and capacitances between the neighbors, but also with transmitter and receiver impedances[4]. The amount of crosstalk can be controlled by controlling the spacing between neighboring wires, or by using each second wire in the bundle as a grounded shield[14]. We have only determined crosstalk by simulations, see the example below.

Power consumption of the wire, finally, is proportional to voltage swing, V_s (here assumed equal to the supply voltage) squared. For a terminated transmission line (far end terminated by Z_0), the input impedance is always Z_0, making its power consumption independent of wire length (assuming random data):

$$P = \frac{V_s^2}{2Z_0} \tag{9}$$

For an open wire, the power consumption is lower and given by[6]:

$$P = \frac{V_s^2}{8Z_0} \quad \text{for } t_d > T/2 \tag{10}$$

$$P = \frac{V_s^2}{4Z_0}\frac{t_d}{T} = \frac{1}{4}BC_wV_s^2 \quad \text{for } t_d < T/2 \tag{11}$$

where T is the bit length, B the data rate (1/T), t_d is the wire delay and C_w=cL is the total wire capacitance. In this case the power consumption is proportional to wire length up to the length at which t_d=T/2, and then constant. Short wires (t_d<T/2) behave just like a capacitor. Normally we use the supply voltage as wire swing, but we could save power by using a reduced swing. Often it is possible to find a power optimal swing, by balancing the wire power consumption with the power consumption in the amplifier used to restore the swing to the logical levels[15,16]. Using reduced swing leads to extra delay due to the delay of this amplifier. As an example, we may save about 2x power at the highest data-rates and with a penalty of about 150ps extra delay[16].

3. A NEW SCHEME FOR GLOBAL INTERCONNECT

We propose a new scheme to global interconnect by combining the knowledge about wires with the idea of Networks on Chip, particularly the SoCBUS concept[9,10]. As demonstrated above, using upper level metals, and possibly improving the fabrication process by slightly increase the metal thickness of these layers, facilitates delays close to what is given by velocity of light. In the same time each wire can carry a considerable data-rate, and can be bundled to very efficient buses. The simplest concept of NoC is based on a 2-dimensional mesh of routers, which then are pair wise interconnected by multiwire links[9,10], see fig. 4. Each router is then attached to an IP-core (through a 5th port). The actual 2D-mesh is mapped to a real layout, keeping the topology but adapting to the actual sizes and positions of the IP-cores. Also I/O and external memory are utilizing the same network through special I/O interfaces and memory management units; see fig. 4. To keep full flexibility at high data-rates, we further demand that the NoC is independent on link delays.

In order to keep latency as small as possible, we will use an upper metal layer with thick metal for the global links. Metal thickness must be sufficient to fulfill the constraint given in eq. (8). Assuming copper, a maximum length of 2cm (corresponding to maximum chip size) and a reasonable aspect ratio of 2 (width/thickness ratio) indicate that we need a metal thickness of about 2µm. This gives an estimated wire resistance rL of 43Ω, compared to $2Z_0\ln 2$=69Ω with Z_0=50Ω. 2µm is somewhat thicker than existing standard thickness in contemporary processes, but has been reported for high performance microprocessors[17]. By keeping the topology from fig. 4, we do not need to cross wires in this layer, so a single layer is sufficient. Still we need an additional layer as ground plane, to assure well-behaved transmission lines. This ground plane can of course simultaneously be utilized either for ground or for supply voltage distribution. Also, the data transport through the router should be as simple as possible in order to keep latency low. For this purpose, the SoCBUS concept[9,10], utilizing circuit-connected data connections without local buffer memories, is excellent.

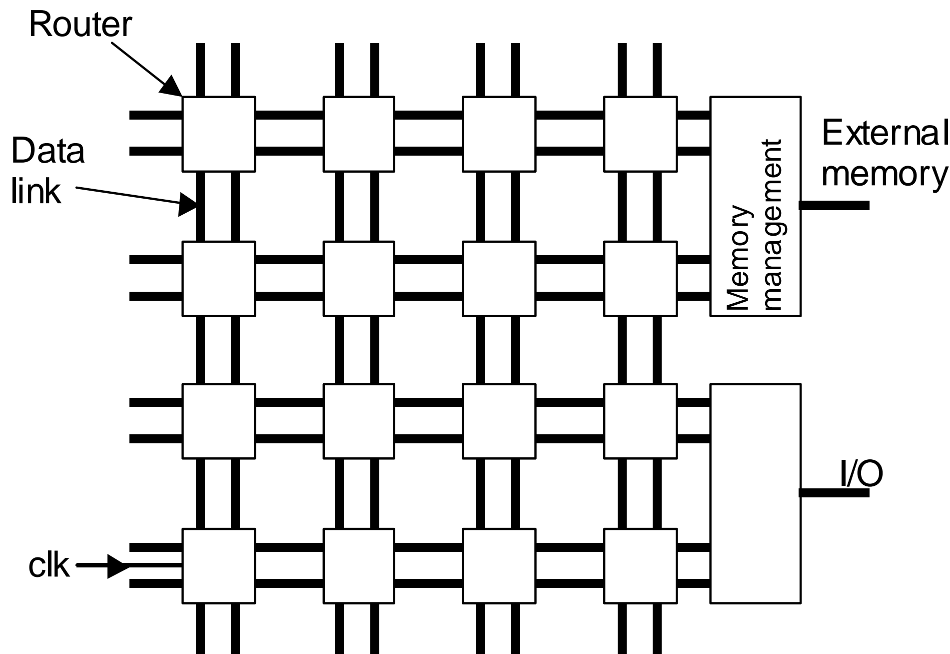

Fig. 4. A 2D mesh structure for a network on chip (NoC). The thick lines are unidirectional links, including a strobe, which also may carry a clock. Each router is assumed to be connected to one IP-core through a 5th port (in a lower metal layer).

In order to maximize data throughput, we should run the wires at their maximum data-rate, still allowing for maximum path lengths. Future chips are most probably limited to a size of the order of 2cm, so the longest link is then less than 2cm making the maximum delay (~$2t_d$) reach 270ps. As this is larger than the bit time (100ps at 10Gb/s), we need some form of data retiming (or handshake) for each data transfer. For the sake of simplicity and low latency, we propose a mesochronous approach (known frequency, unknown phase), according to the following. Let each (unidirectional) data link include a strobe signal in addition to the data bus. The strobe signal is generated in the same way as the data and transported along the data, making it delay-matched to the data. It has a frequency of half the clock frequency and is half a clock period out of phase compared to data. On the receiver side the strobe is used to retime the incoming data to the local clock. Retiming could for example be performed through self-timed self-synchronization[18]. The distributed strobe signals may also be utilized as clock distribution in the NoC subsystem, by choosing the frequency doubled strobe signal from one of the ports of each router as its local clock.

4. A NOC EXAMPLE

As a design example we choose to study a 2cm long copper wire having thickness and width of 2µm and 4µm, respectively. We place a ground plane 3.55µm below the wire to make the wire Z_0 equal to 50Ω. To drive the wire, we use a low impedance inverter driver, fabricated in a 0.18µm process. We tuned the transistor sizes of the driver to W_p=194µm (PMOS width) and W_n=88µm (NMOS width) to achieve the same switching step response as for an ideal driver having its output impedance close to 20Ω. This corresponds to a driver utilizing pre-emphasis (overdrive).

Fig. 5 shows the Hspice step response of our single isolated transmission line in three cases. The upper case uses a low impedance driver, about 20Ω impedance, and an open far end. It gives rise to a limited overshoot, but is very steep, thus allowing very high data-rate. The middle case use 50Ω impedance of the driver and an open end. The lower case, finally, uses 50Ω terminations at both near and far end of the wire. Here we see a large amplitude loss, 50% due to the fact that the line is terminated, and an additional 30% due to wire series resistance. In reality, the terminated case will keep the long skin effect induced tail, seen in fig. 3, but this is not correctly modeled by HSPICE. We will use the upper

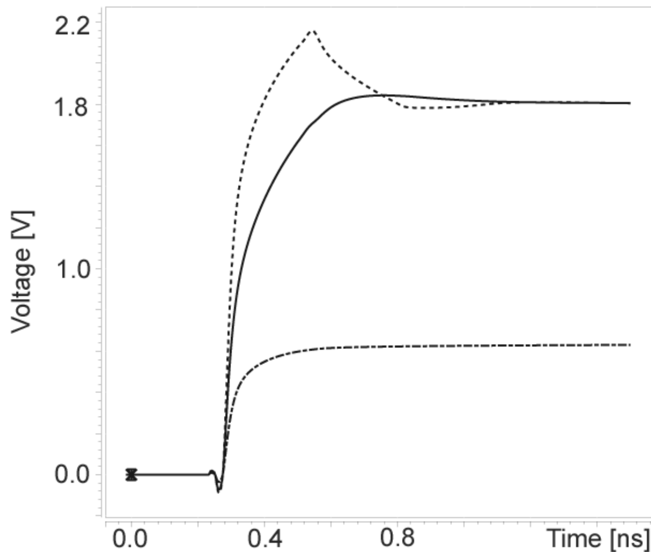

Fig. 5. Hspice step response of a transmission line with cross section 2x4μm^2, length 2cm, Z_0=50Ω, driven by an inverter in 0.18μm CMOS. Pre-emphasis driver and unterminated line (top curve), matched driver and unterminated line (middle curve), matched driver and terminated line (bottom curve).

case, a low impedance driver (20Ω) and an open far end, in order to maximize speed and minimize power consumption. The wire delay was simulated to 200ps, corresponding to a propagation velocity of 10^8m/s. The maximum data-rate was estimated from both Matlab-modeling of the full wire equations (with skin effect) and from HSPICE simulations to 11 and 10Gb/s respectively, where we used the criterion 64% eye opening from the step response. In order to estimate the wire spacing, we simulated the worst-case crosstalk in HSPICE with two neighboring wires, both driven by a fast step using the same driver as for the main wire. Allowing a maximum crosstalk amplitude of 18% (corresponding to an eye-opening of 46% including both dispersion and crosstalk), we conclude that the spacing needs to be 12μm. We therefore need a width per wire of 16μm, making a 16 bit bus with one strobe 272μm wide. Two such buses are utilized for each link (between each pair of routers), facilitating a total bi-directional bandwidth of 320Gb/s at a bus width of 544μm.

A reasonable size of the NoC network is 8x8 routers, leading to a maximum bi-directional, bi-section bandwidth of 2560Gb/s across the chip (2 directions, 8 links, 16 wires per direction per link and 10Gb/s per wire). Assuming that one side of the chip is used as memory interface (supported by a memory management unit) means that the maximum external memory bandwidth supported by Socbus also is 2560Gb/s (which of course is far to high to be managed with the memory bus technique of today). The maximum bidirectional bandwidth available to the IP cores is 20480Gb/s (2 directions, 16 wires per direction for each of 8^2 IP-cores, 10Gb/s per wire). In practice, all networks have a limited ability to utilize the full bandwidth, due to congestion. Preliminary simulations indicate that Socbus, for example, can accept a sustained bandwidth of the order of 10% of the total bandwidth for random traffic[10]. The useful bandwidth in our example is therefore reduced to about 2048Gb/s, still more than 2Tb/s. We may note that surplus bandwidth is quite useful for mitigating the effect of congestion in networks with high traffic.

The total power consumption of this network can be estimated as follows. As the average link length is about $L_{av}=L_c/8$ in this case, where L_c is the chip edge, we will use eq. (11) for this estimation. We have a total link length of 2*2*2=8cm (2 directions, 2 dimensions, 2cm total length per dimension) and each link consists of 16 wires. This yields a total wire length of 128cm, corresponding to a total electrical length (t_d) of about 8.5ns. With T=100ps, V_s=1.8V (0.18μm process) and Z_0=50 we get P_{tot}=1.38W. Using a smaller voltage swing can further reduce this power consumption; for the highest speed a reduction of 2x may be realistic in the actual process. Further reductions can be foreseen in future processes (because of smaller supply voltage and higher speed, allowing higher amplifier gain).

It may also be interesting to investigate area utilization of the metal layer used. With the above example, and assuming a chip size of 10 mm by 10 mm, we may estimate the average chip area available per router and IP core to $(10/8)^2$ mm^2. Assuming a link width of 0.544 mm as above and a router area of 0.544^2 mm^2, we may estimate the area used by the

links to 4*0.544*((10/8-0.544)/2) mm^2, corresponding to 45% of the total area per router and IP core, see fig. 6. This indicates quite a large utilization of the metal layer used for the links, meaning that we cannot deviate too much from a geometrically regular mesh structure. However, for larger chips situation improves. For a 20 mm by 20 mm chip, for example, the metal layer utilization is reduced to 31%, allowing considerably less regular mesh structure.

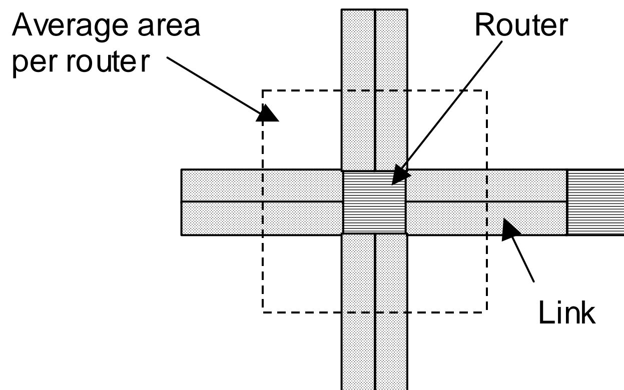

Fig. 6. Schematic layout of links around a router. Note that the 5th link to the IP-core is assumed to use a lower metal layer.

V. CONCLUSION

We investigated the properties of global interconnect, using a microstrip model of the wire, concerning key performances as data delay, maximum data-rate, crosstalk and power consumption. The result from this analysis shows, that by utilizing an upper, thicker metal layer, of about 2µm thickness, we may reach delays comparable to velocity-of-light delay. In the same time it is possible to reach global data-rates up to 10Gb/s per wire (using a 0.18µm process). Assuring large enough spacing between neighboring wires can mitigate crosstalk, and using reduced voltage swing may reduce power consumption. Based on these findings, we propose a new scheme for global interconnect, based on the utilization of microstrip lines utilizing two upper-level metal layers, one for wires and one for ground plane, combined with a 2D network-on-chip mesh concept. We demonstrate an example of an 8 by 8 2D mesh of routers, each serving one IP-core, and utilizing very simple routers of the SoCBUS type. We show that such a network-on-chip can offer a total bandwidth of more than 20Tb/s to the 64 IP-cores and has a bi-directional bisection bandwidth of more than 2Tb/s. The bandwidth 2Tb/s is also what can be offered to serve each chip edge with I/O capacity, if a future I/O can manage such a bandwidth (It is not impossible though[6]). The total power consumption related to the wires is estimated to less than 1.4W in this example.

REFERENCES

1. C. Mead and L. Conway, *Introduction to VLSI Systems*, Addison-Wesley, 1980.
2. H. Bakoglu, *Circuits, Interconnection and Packaging for VLSI*, Addison-Wesley, 1990.
3. C. Svensson and M. Afghahi, “On RC Line Delays and Scaling in VLSI Systems”, *Electronic Lett.*, **24**, 562-563, 1988.
4. W.J. Dally and J. W. Poulton, *Digital Systems Engineering,* Cambridge University Press, 1998.
5. A. Deutsch, et. al., ”On-Chip Wiring Design Challanges for Gigahertz Operation”, *Proc. IEEE*, **89**, 529-555, 2001.
6. C. Svensson, “Electrical Interconnects Revitalized”, *IEEE Trans. of VLSI Systems*, **10**, 777-788, 2002.
7. L Benini and G. De Micheli, “Network on Chips: A New Soc Paradigm”, *IEEE Computer*, **35**, 70-80, 2002.
8. K. Goossens, J. Dielissen, J. van Meerbergen, P. Poplavko, A. Radulescu, E. Rijpkema, E. Waterlander, and P. Wielage, "Guaranteeing the quality of services in networks on chip", *Networks on Chip*, Axel Jantsch and Hannu Tenhunen, Kluwer, 2003.
9. Dake Liu, Daniel Wiklund, Erik Svensson, Olle Seger, and Sumant Sathe, “SoCBUS: The solution of high communication bandwidth on chip and short TTM” *Proc of the Real-Time and Embedded Computing Conference*, Gothenburg, Sweden, 2002.

10. Daniel Wiklund and Dake Liu "SoCBUS: Switched Network on Chip for Hard Real Time Systems" to appear at *the IPDPS'03*, Nice, France, 2003
11. C. Svensson and G. Dermer, "Time Domain Modeling of Lossy Interconnects", *IEEE Transactions on Advanced Packaging*, **24**, 19, 2001.
12. J. M. Rabaey, A. Chandrakasan and B. Nikolic, *Digital Integrated Circuits*, Second edition, Prentice Hall, 2003.
13. A. Fiedler, R. Mactaggart, J. Welch and S. Krishnan, "A 1.0625Gbps Transceiver with 2x-Oversampling and Transmit Signal Pre-emphasis", *1997 Int. Solid State Circuit Conference, Digest of Technical Papers*, 238-239, 1997.
14. Oh-Kyong Kwon, R. Fabian W. Pease, "Closely Packed Microstrip Lines as Very High-Speed Chip-to-Chip Interconnects", *IEEE Transactions on Components, Hybrids, and Manufacturing Technology*, **CHMT-10**, 1987.
15. C. Svensson, "Optimum Voltage Swing on On-Chip and Off-Chip Interconnects", *IEEE J. Solid-State Circuits*, **36**, 1108-1112, 2001.
16. P. Caputa and C. Svensson, "Low-Power, Low-Latency Global Interconnect", *15th Annual IEEE International ASIC/SOC Conference, Rochester*, Sept. 25-28, 2002.
17. B. J. Benschneider et. al., "A 1GHz Alpha Microprocessor", *2000 IEEE Solid-State Circuits Conference , Digest of Technical Papers*, 86-87, 2000.
18. F. Mu and C. Svensson, "Self-tested Self-Synchronization Circuit for Mesochronous Clocking", *IEEE Trans. on Circuits and Systems – II*, **48**, 129-140, 2001.