RF Interconnects for Communications On-chip[1]

M.-C. Frank Chang, Eran Socher, Sai-Wang Tam
Electrical Engineering Dept. UCLA
Los Angeles, CA 90095
001-1-310-794-1633

{mfchang,socher,roccotam}@ee.ucla.edu

Jason Cong, Glenn Reinman
Computer Science Dept. UCLA
Los Angeles, CA 90095
001-1-310-206-2775

{cong,reinman}@cs.ucla.edu

ABSTRACT
In this paper, we propose a new way of implementing on-chip global interconnect that would meet stringent challenges of core-to-core communications in latency, data rate, and re-configurability for future chip-microprocessors (CMP) with efficient area and energy overheads. We discuss the limitation of traditional RC-limited interconnects and possible benefits of multi-band RF-interconnect (RF-I) through on-chip differential transmission lines. The physical implementation of RF-I and its projected performance versus overhead as the function of CMOS technology scaling are discussed as well.

Categories and Subject Descriptors
B.7.1 [**Integrated Circuits**]: Types and Design Styles – *Advanced technologies, Microprocessors and microcomputers, VLSI.*

General Terms
Performance, Design.

Keywords
RF-Interconnect, Network-on-Chip, Chip MultiProcessors.

1. INTRODUCTION
The CMOS technology scaling trend has been driven by the desire for higher performance and lower cost integrated circuits for several decades. Scaling enabled the decrease of digital gate capacitive loads, thereby reducing gate latency and also reduced the silicon area of digital gates, enabling the integration of more logic in the same chip area. Both trends contributed to microprocessor performance in terms of instructions performed per second, since lower latency enabled higher clock frequency, and more logic enabled more parallelism.

Two physical facts present significant challenges to the continued performance improvement via simple device scaling. One is dissipated power. As clock frequency increases, more power is consumed, up to a point where it causes serious thermal and reliability issues. Increased leakage, as a result of scaling, further cuts into the power budget, thus limiting the increase of frequency even more. The second physical fact is the scaling effect on the properties of metal wires that serve as device and unit interconnects. Previously regarded negligible, wire contribution to latency and power consumption became significant with scaling due to increased resistance and more significant capacitance compared with MOS devices. While local wires still scale nicely since their length is getting shorter, global wires used for communication across the chip are a potential inhibition to improving microprocessor performance.

One of the current trends to meeting the need for improved microprocessor performance with tolerable power consumption is chip multi processors (CMPs) [1]. Under this approach, scaling is used to reduce the area of processor cores while keeping the clock operation frequency relatively constant. Integrating more cores on the same chip area allows a performance benefit through parallelization. Communication between on-chip cores and shared on-chip caches can be done using the traditional bus. As the number of cores and caches banks further increases, the shared bus does not scale so easily, resulting in more latency for communication and consuming more power and area. An alternative architecture that seems to scale more easily is dubbed network-on-chip (NoC), in which routers are used in conjunction with every on-chip core and cache bank. In a simple mesh topology, each router can communicate with his four orthogonal neighbors and with his adjacent core or cache. Multi-bit repeated wire based buses are still assumed to connect the mesh-network routers. As a result, although close neighbors can easily communicate with high bandwidth and low latency in an NoC, communication between distant points across the chip incurs longer latencies. The bottleneck gets even more severe when overall chip communication is considered, when larger traffic increases router congestion and makes actual latencies even longer.

2. RF-I Concept
Repeated wires have limited speed both in terms of latency and in terms of achievable data rate due to the high resistance-capacitance (RC) product induced by the wires themselves. This product is limiting the speed since the signaling in this approach is based on charging and discharging of the wire segments to send information. The result is a poor use of the frequency bandwidth offered by super-scaled CMOS devices [2][3][4]. In our proposed RF-I concept, transmission of data is achieved by modulating an electromagnetic (EM) wave that is guided along a wire that serves as a transmission line on-chip. Since the EM wave travels at the effective speed of light in the dielectric layers of the chip, latency is the smallest that can physically be achieved. The available data rate also benefits in two ways. Baseband data rate is limited by the high attenuation of high frequencies compared with DC due to wire resistance. As higher data rates are desired with scaling, equalizing the loss dispersion from DC to higher frequencies becomes more difficult.
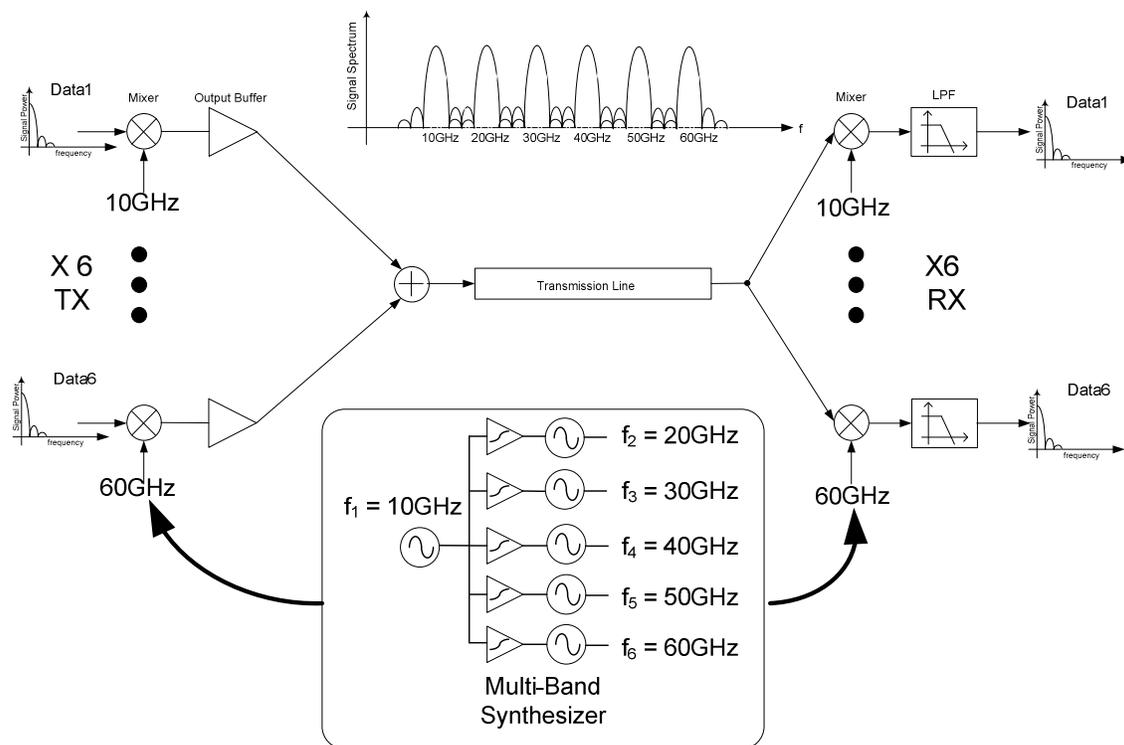
Figure 1. FDMA RF-I concept illustration

Modulating a high frequency carrier on a transmission line offers relatively less variation in attenuation even at high modulation rates (since the data bandwidth to carrier frequency ratio $\Delta f/f_c$ is much small than at baseband), thus simplifying the required equalization, saving power and area. Another important advantage is the ability to modulate several high frequency carriers, similar to the FDMA concept in wireless communication, taking advantage of over a 100GHz of available on-chip bandwidth to achieve unprecedented aggregate data rates on-chip [3].

Using FDMA for RF-I introduces an even wider range of architectural opportunities if multiple access points are used with the same transmission line (TL). Transmissions are then not limited to just fixed point-to-point communication, but frequencies can be dynamically assigned to different transmitting and receiving units, all sharing the same TL. The total aggregate data rate possible on a given TL can then be divided between several users at different access points according to the current communication bandwidth needed at a certain time. This would lower the area and power overhead required to support worst case data rates for each case individually. Tuning several receivers to the same transmitting

frequency would enable parallel broadcasting of data or control messages to several units, saving time and energy.
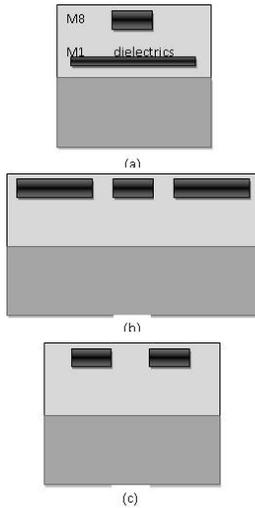
3. CMOS Transmission Lines

Super-scaled CMOS processes offer interesting possibilities for on-chip transmission lines for guiding EM waves at mm-wave frequencies. Since scaling of CMOS devices requires more wiring, current processes offer more layers of metallization, enabling the realization of transmission lines at larger distance from the lossy silicon substrate. Modern CMOS processes typically offer 6 fine pitch metal layers and additional two or more thicker metal layers. This metal and dielectric stack allows the implementation of various on-chip transmission lines, with ever increasing distance from the lossy silicon substrate. Transmission lines have been design and implemented in CMOS for various circuit applications, mainly as passive resonant structures and as matching and filtering devices [5]. Design consideration for transmission lines used as a high data rate on-chip communication media are different. Several transmission line structures are possible for implementation. The first is a microstrip, in which the signal line is implemented in a thick, top metal layer and ground plane can be either the silicon substrate or a low metal layer. Using a low metal level as ground may shield the EM waves from the lossy substrate, but typically add to the resistive loss in the signal line due to the decreased characteristic impedance. The structure is limited in terms of immunity to interference. Since high data rate are desired, we should consider the implementation of several adjacent transmission lines for parallel transmission of multi-bit data. In order to achieve a high density of data rate per width of the transmission line bus, spacing between transmission lines has to

be minimized, which makes interaction between adjacent microstrips dominant compared with the signal-ground mode, making crosstalk between neighboring signals significant.

Figure 2. Transmission line (TL) structures in CMOS: (a) microstrip TL, (b) coplanar waveguide TL, and (c) differential TL

Another popular structure for on-chip transmission lines is the co-planar waveguide. In this structure the ground plane is typically implemented using two lines on each side of the signal line, all in the top metal layer. Fields are then confined to the top part of the metal stack, away from the lossy substrate. Due to the shorter signal-ground distance, the characteristic impedance tends to be lower than microstrip and the resistive loss tends to be higher. The ground planes can be effective in shielding between adjacent transmission lines, but their width tends to increase the available pitch of the transmission line bus, thus lowering the achievable data rate density.

In our work we chose to concentrate on a differential transmission line, which combines the advantages of the above two structures. In this structure two identical signal lines are implemented in proximity in top metal layers and used to guide a differential mode excited by a differential signal. Confinement of the fields between the two lines distances them from both the substrate and adjacent transmission lines. Thus, both substrate loss and crosstalk (or inter-channel interference) can be minimized.

Setting the dimensions of the selected TL is an intricate tradeoff between several performance metrics, which include loss, latency, crosstalk between adjacent TLs and the frequency dependence of the loss and latency. Loss, which typically increases with frequency, is important since it determines the signal level at the receiver. Higher loss per mm would limit either the communication distance or the maximum carrier frequency that can be used. Alternatively, it would force an increase in power consumption and area required by the transmitter, receiver or both to compensate in gain for the loss in signal level. Narrowing of TL signal line width tends to increase the resistive loss. Widening the spacing between the two differential lines and the lines themselves tends to increase the loss in the silicon substrate.

Latency in the TL is important since it is one of the motivations for using TLs instead of repeated wire bus. The latency per mm is determined by the effective speed of light of the EM waves guided by the TL, which is set by the effective dielectric constant of the EM wave mode. Thus it would be determined by the dielectric layers that most of the electric field flows through. For narrow and close differential TLs, the dielectric constant would be close to that of the top dielectric layers (with a relative dielectric constant of about 4). Wider and more spaced lines would include more of the substrate in the field path, thus increasing the effective relative dielectric constant toward that of silicon, which is 11.7, and thus increasing the TL latency by about 50%.

Differently from transmission lines used for off-chip connection, the characteristic impedance of RF-I transmission lines does not have to match an existing standard (such as 50 or 75Ω). However, transmission line characteristics such as impedance and phase velocity need to be uniform enough with frequency to minimize signal distortion effects that may reduce the available data rate and increase BER, especially due to the very broadband nature of RF-I. Most of our designs used a 3μm wide differential TL with a spacing of 3μm. The latency of such TLs is about 70 ps/cm and the loss is 15 dB/cm, both at 60 GHz

## 4. On-Chip Multiple Carrier Generation

For FDMA RF-I to materialize, multiple carrier frequencies have to be generated on-chip in the mm-wave range. Furthermore, this generation has to be efficient in terms of power consumption and silicon area for the concept to be viable. Current solution to on-chip mm-wave frequency generation include phase-lock-loops (PLLs), which consume significant power, especially in the high frequency divider chain. Implementation of multiple frequency generators on-chip using independent PLLs would consume prohibitive power and silicon area to be practical to RF-I. In a recent work [6] we proposed the use of sub-harmonic injection locking for concurrent generation of multiple mm-wave frequency carriers on-chip by locking them simultaneously to a single reference. The principle was demonstrated in a 30GHz and a 50GHz VCOs that were locked using a 10GHz source utilizing an internally generated 3rd and 5th harmonics of the source for injection locking. Each VCO consumes 4mW of power and locking ranges of up to 5.2GHz were demonstrated, depending on the harmonic that was used to lock.
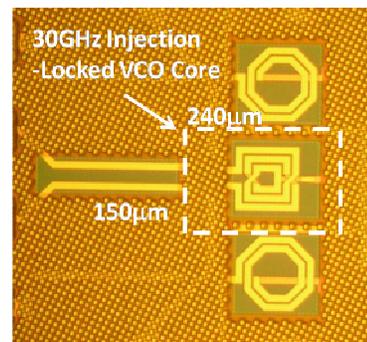
Figure 3. Die micrograph of an injection locked sub-harmonic frequency generator showing on-chip top metal inductors.

5. On-Chip RF-I Implementation

As a design example we can consider the RF-I of which the layout is shown in Fig.4a in IBM 90nm CMOS process. The link is based on BPSK modulation of a 20GHz carrier wave. The 20GHz differential source signal is buffered and then phase modulated using a Gilbert cell and a differential data stream at 5Gbps. The resulting phase modulated wave is then transmitted along a 1cm differential TL. The carrier wave itself is also transmitted on a second differential TL to serve as source synchronization on the receiver side. On the receiving end, we find a downconverting mixer that multiplies the transmitted modulated signal with the transmitted and buffered carrier wave. The baseband result is then fed into a data recovery differential stage with capacitive boosting of high frequencies. Most of the area consumed by the circuit is used for implementing the passive devices. The 1cm long transmission line uses the top two metal layers and covers 0.12 mm^2, including the required spacing. Three inductors are used in the transmitter, two in buffers and one in the upconversion mixer, all centered for operation around 20GHz and each of them with an area of 0.012 mm^2, almost exclusively using the top metal layer. Another similar inductor is used in a receiver buffer. Metal-insulator-metal (MIM) capacitors, used for coupling and loading and use mostly top metal layer, contribute about 2400 μm^2 to the transmitter area and 1500 μm^2 to the receiver area. Devices that require use of actual silicon and gate area (such as transistors and resistors) consume an area of 1600 μm^2 in the transmitter and almost 2000 μm^2 in the receiver.

(a)

(b)
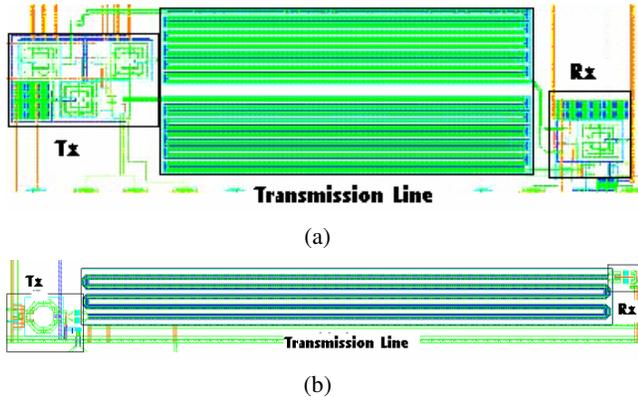
Figure 4. Layouts of a BPSK implementation of a 20GHz RF-I (a) and an ASK implementation of a 50GHz RF-I (b)

Using BPSK-based RF-I has its limitations as synchronization is required. Not only is the carrier frequency of the transmitter required on the receiver side but the phase has to be matched as well. This requirement call for source synchronization and/or source recovery circuits that add to the area and power overhead of the RF-I link. An alternative to using phase shift keying is to use amplitude shift keying (ASK), the layout of which is shown in Fig. 4b. In this approach amplitude modulation is used in the transmitter, while simple envelop detection is used to demodulate the data on the receiver side. The carrier wave power can be simply transfers or not to the transmission line, and the modulated power can be measured at the receiving end. Filtering is required at the receiver to separate the power received from different carrier frequencies. However, this

filtering can be part of the resonant coupling structure between the TL and the receiver.

6. RF-I Expected Performance with Scaling

When evaluating RF-I performance, it has to be compared with current and future alternatives. The current solution to on-chip communication is the repeated wire, or its multi-bit extension. In this approach global wiring on the chip is implemented in the top metal layers. Since the resistance and capacitance product of these wires induces significant latency and unacceptable rise and fall times when mm-length wires are considered. The use of optimized repeaters changes the latency dependence upon length from quadratic to linear and allows semi-global data transmission under one clock cycle. Since the data rate on every wire is limited by RC time constants, a wide bus is required to achieve a wide data bandwidth. Scaling allows a higher density of wires in the bus, so that when more repeaters are used a higher data rate density can be achieved. In this study, the delay of a global repeated wire was optimized in circuit simulation for 90nm CMOS technology. Large uniform-size repeaters of 300x at a distance of 1mm were chosen for delay minimization (The performance of non-uniform sized repeaters and wires can also be easily estimated using the IPEM models described in [8]). Estimation of delay scaling was done using ITRS projections for global wire resistance and capacitance, assuming:

$$T_{delay,projected} = T_{delay,90nm} \sqrt{\frac{T_{d,global\,wire,projected} FO4_{projected}}{T_{d,global\,wire,90nm} FO4_{90nm}}}$$

The consumed power in this traditional approach can be normalized to a figure of merit of the energy consumed per bit transmitted. This energy is significant since the whole length of the wire capacitance is charged and discharged in transmission, in addition to the input and output capacitance of the repeaters. The energy consumption of optimal repeated wires in 90nm technology was also simulated. Capacitance of other nodes was estimated using ITRS projections for global wire capacitance and repeater contribution was estimated assuming a similar effective resistance as in 90nm and decreasing repeater spacing using its proportionality to $\sqrt{(FO4/T_{d,global\,wire})}$. Both capacitance and the supply voltage are not expected to decrease much with scaling [7]. Even though repeater capacitance is expected to decrease with scaling, the effect is largely compensated by the larger number of repeaters required to keep the same data rate. As a result, the energy efficiency is not expected to improve much with scaling. Since it is projected that there will be close to 800,000 repeaters in 70nm designs [9], the energy consumption of the traditional RC bus is a significant challenge.

A possible alternative to global on-chip interconnects that was proposed are optical interconnects [10]. In this approach, light is modulated using the data and transmitted using on-chip waveguides to be demodulated on the receiver side. Using the projected performance in [10], it is possible to extrapolate the achievable data rate, latency, area and consumed energy per bit for on-chip global communication. The estimated latency (which includes propagation of 10.45ps/mm, driver, modulator, detector and amplifier delays as detailed in [10]) is similar to that of RF-I, since the effective speed of light is similar. The expected data rate is not as high, since the integration of multi-wavelength transmitters is not trivial. The energy per bit is estimated using

the total 1.4 W assumed for a network achieving an aggregate data rate of 1344 Gbps. Data rate density is expected to improve as more wavelengths are introduced, allowing higher data rates per waveguide. When scaling is considered, it is important to note that since at least some of the devices used in this approach are not CMOS devices, the CMOS scaling trend does not make their performance better. Improvement is therefore not expected to grow at the same rate.

Contrary to standard voltage signaling in global CMOS interconnects, RF-I performance is expected to improve with scaling. The most apparent benefit of scaling is the transistor speed and size. As predicted by ITRS [7], CMOS transistors speed, characterized by their f_T is expected to increase. As a result, these transistors would enable the use of higher operation frequencies. One result of the frequency increase is the ability to generate and modulate higher carrier frequencies and transmit them on the same transmission lines. Oscillation frequency of 324GHz was recently demonstrated in 90nm CMOS [11]. Another result is the ability to modulate the same carrier frequency at a higher data rate. Both results allow the transmission of a higher aggregate data rate on the same transmission line. Scaling reduces the area of transistors as well, making the active area required by RF-I circuits smaller.

Table 1. CMOS oscillation speed scaling

RF CMOS vs. Tech	90nm	65nm	45nm	32nm	22nm	16nm
f_T (GHz)	120	170	240	320	400	490
f_{max} (GHz)	200	270	370	480	590	710
Max RF carrier	324	432	592	768	944	1136
Max Aggregate Data	160	216	296	384	472	568

The increase in carrier frequencies benefits area as well. Since most of the area is consumed by passive elements and inductors in particular, inductor area is of special concern. However, inductor area scales down with the intended operation frequency. The reason is that circuits such as oscillators and tuned amplifiers the important characteristic of the inductor is the imaginary part of its impedance, proportional to both the inductance and the frequency. Thus, keeping the required impedance unchanged means a lower inductance for a higher operation frequency. Since the inductance is proportional to the winding total length, the lower inductance would fit into a smaller area. Moreover, the skin effect limits the current flowing in the inductor cross section to smaller areas as frequency increases, thus enabling the use of narrower windings in higher frequency inductors and reducing the inductor area even further.

In the RF-I approach, EM waves are generated, modulated, transmitted and demodulated continuously. As a result, the area and power consumed are fixed according to the number of carriers used. Therefore, as both carrier frequencies and modulation rates increase, the area and energy consumed per bit sent are constantly reduced. Accordingly, each new technology node would increase the aggregate data rate on the one hand, and decrease the normalized area and energy consumption on the other hand. Table 2 shows the scaling trend of performance with the technology node. The increase in the number of carrier frequencies and the modulation speed are based on the ITRS

transistor speed trend of Table 1. The power is estimated using the power consumed in the design of a single frequency RF-I in 90nm CMOS, averaged in the 10-60GHz range. The total power is assumed to be proportional to the number of carrier frequencies. When scaling is considered, we assumed the same power per channel as in 90nm technology. While higher frequency carriers added may consume more power, the lower power consumed by lower frequencies due to higher device RF gain and lower device parasitics (or higher f_{max} of the CMOS device) is expected to compensate that on average.
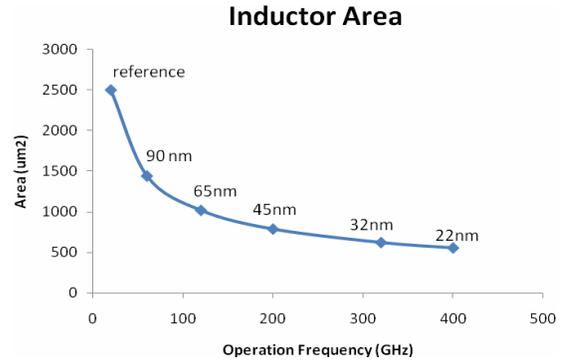
Figure 4. Inductor area scaling with operation frequency

Table 2: Scaling trend of RF-I

Technology	# of Carriers	Data rate per band (Gb/s)	Total data rate per wire (Gb/s)	Power (mW)	Energy per bit(pJ)	Area (Tx+Rx) mm^2
90nm	6	5	30	36	1.2	0.0107
65nm	8	6	48	48	1	0.0112
45nm	10	7	70	60	0.85	0.0115
32nm	12	8	96	72	0.75	0.0119
22nm	14	10	140	84	0.6	0.0123

It is therefore possible to compare the performance trend with scaling of the three approaches for future interconnects discussed, shown in Fig. 5. While the traditional bus approach has increased latency with scaling, RF and optical interconnects have similar latency that remains fairly constant with scaling, enabling global transmission in less than a clock cycle. Both RF and optical interconnects have a significant energy consumption benefit compared with the traditional bus, but RF-I scales better and eventually shows a benefit over optical. Data rate density is expected to improve in all three technologies, where the bus benefits from wire pitch and RF benefits from the number of carrier frequency and modulation speed possible. RF-I has the advantage of using the standard digital CMOS technology, while optical-I requires integration with on-chip and off-chip non-CMOS devices adding to package complexity and cost.

7. RF-I Impact on NoC CMP Performance

In a recent paper [12], the possible effect of RF-I on the performance of NoC CMPs was studied. As an NoC test case, a 10×10 mesh that includes 64 cores, 32 L2 cache banks and 4 memory access nodes in 32nm CMOS was chosen. RF-I was added to the mesh network to connect routers in the center of the mesh and the centers of its four quadrants, thus decreasing the physical latency of such transmissions and allowing concurrent communication on the same transmission lines using different frequencies. Assuming a 400mm^2 die, the study demonstrated that in exchange for 0.13% of area overhead on the active layer, RF-I can provide an average 13% (max 18%) boost in application performance, corresponding to an average 22% (max 24%) reduction in packet latency based on detailed, cycle-accurate architecture simulation using the multi-core simulator MC-Sim developed at UCLA on a set of SPLASH benchmark examples.
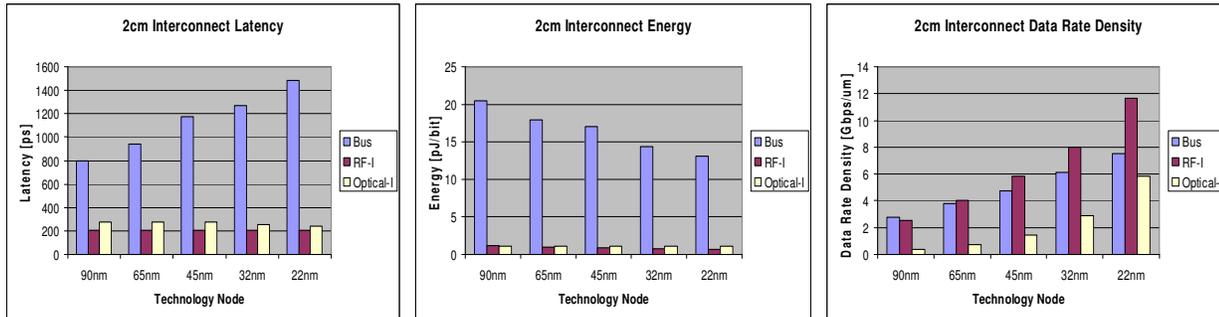

Figure 5. Interconnect technology comparison for a global 2cm on-chip distance of latency, energy consumption per bit

8. Summary and Ongoing Work

A new approach to global interconnect is presented, useful for future design of CMPs that employ on-chip networks. The approach is based on FDMA RF circuits and on-chip differential transmission lines, all implemented in standard CMOS. The RF-I approach benefits from CMOS scaling and is therefore expected to keep improving as demand for on- and off-chip bandwidth increases. In addition, it provides significant advantages in terms of latency, power, and reconfigurability.

RF-interconnects also introduce many interesting and important physical design and architectural design problems. For example, when we overlay an RF-interconnect bus on a mesh-based NoC network, we are effectively adding a set of "short-cuts" to the network, where the total number of such short-cuts is limited by the overall RF-interconnect bandwidth. The optimal addition of these short-cuts is an interesting optimization problem. Moreover, given the reconfigurability of RF-interconnects, we may change the locations and/or bandwidth allocation of these short-cuts either during compilation time or at runtime over different phases of an application. These present interesting architectural design problems. Finally, physical implementation of chip designs that include both processing core and RF circuits for on-chip communications is a challenging mixed-signal physical design problem. We are actively working in these directions.

9. REFERENCES

[1] L. Benini, G. De Micheli, "Networks on Chips: A New SoC Paradigm," *IEEE Computer*, vol. 35, no. 1, Jan. 2002.

[2] E. Socher and M. F. Chang, "Can RF Help CMOS Processors?", *Invited Paper*, IEEE Communication Magazine, Vol. 45, No. 8, pp. 104-111, 2007.

[3] J. Ko, J. Kim, Z. Xu, Q. Gu, C. Chien, and M.F. Chang, "An RF/Baseband FDMA-Interconnect Transceiver for Reconfigurable Multiple Access Chip-to-Chip Communication," in *2005 IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, February 2005.

[4] Q. Gu, Z. Xu, J. Ko and M.F. Chang, "Two 10Gbps/pin Low Power Interconnect Methods for 3D IC", in *2007 IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, February 2007.

[5] D.Q. Huang, W. Hunt, N.Y. Wang, T. W. Ku, Q. Gu, R. Wong, and M. F. Chang, "A 60GHz CMOS VCO Using On-Chip Resonator with Embedded Artificial Dielectric for Size, Loss and Noise Reduction," in *2006 IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, February, 2006.

[6] S.-W. Tam, E. Socher, A. Wong, Y. Wang, L. Vu and M. F. Chang, "Simultaneous sub-harmonic injection-locked mm-wave frequency generators for multi-band communication in CMOS", in *2008 IEEE Radio Frequency Integrated Circuits (RFIC) Digest of Technical Papers*, June 2008.

[7] International Technology Roadmap for Semiconductors: Semiconductor Industry Association, 2006, www.itrs.net

[8] J. Cong and Z. (D.) Pan, "Interconnect Performance Estimation Models for Design Planning" *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, June 2001, vol. 20, no. 6, pp. 739-752.

[9] J. Cong, "An Interconnect-Centric Design Flow for Nanometer Technologies," *Proc. of the IEEE*, April 2001, vol. 89, no. 4, pp 505-528.

[10] N. Kirman, M. Kirman, R.K. Dokania, J.F. Martinez, A.B. Apsel, M.A. Watkins, and D.H. Albonesi, "Leveraging Optical Technology in Future Bus-based Chip Multiprocessors," In *Proceedings of MICRO-39*, December 2006.

[11] D. Huang, T. R. LaRocca, L. Samoska, A. Fung and M.F. Chang, 324GHz CMOS Frequency Generator Using Linear Superposition Technique, in *2008 IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, February, 2008.

[12] M. F. Chang, J. Cong, A. Kaplan, M. Naik, G. Reinman, E. Socher and S.-W. Tam, *CMP Network-on-Chip Overlaid With Multi-Band RF-Interconnect*, IEEE High Performance Computer Architecture (HPCA '08).