The Case for Hybrid Photonic Plasmonic Interconnects (HyPPIs): Low-Latency Energy-and-Area-Efficient On-Chip Interconnects

Volume 7, Number 6, December 2015

Shuai Sun
Abdel-Hameed A. Badawy
V. Narayana
Tarek El-Ghazawi
Volker J. Sorger

Chip-Scale Interconnect Performance

DOI: 10.1109/JPHOT.2015.2496357
1943-0655 © 2015 IEEE
The Case for Hybrid Photonic Plasmonic Interconnects (HyPPIs): Low-Latency Energy-and-Area-Efficient On-Chip Interconnects

Shuai Sun,1 Abdel-Hameed A. Badawy,1,2 V. Narayana,1 Tarek El-Ghazawi,1 and Volker J. Sorger1

1Department of Electrical and Computer Engineering, The George Washington University, Washington, DC 20052 USA
2Department of Electrical Engineering, Arkansas Tech University, Russellville, AR 72081 USA

DOI: 10.1109/JPHOT.2015.2496357
1943-0655 © 2015 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Abstract: Moore’s law for traditional electric integrated circuits is facing increasingly more challenges in both physics and economics. Among those challenges is the fact that the bandwidth per compute on the chip is dropping, whereas the energy needed for data movement keeps rising. We benchmark various interconnect technologies, including electrical, photonic, and plasmonic options. We contrast them with hybrid photonic–plasmonic interconnect(s) (HyPPI(s)), where we consider plasmonics for active manipulation devices and photonics for passive propagation integrated circuit elements and further propose another novel hybrid link that utilizes an on-chip laser for intrinsic modulation, thus bypassing electrooptic modulation. Our analysis shows that such hybridization will overcome the shortcomings of both pure photonic and plasmonic links. Furthermore, it shows superiority in a variety of performance parameters such as point-to-point latency, energy efficiency, throughput, energy delay product, crosstalk coupling length, and bit flow density, which is a new metric that we defined to reveal the tradeoff between the footprint and performance. Our proposed HyPPIs show significantly superior performance compared with other links.

Index Terms: Optical interconnect, plasmonics, hybridization, hybrid photonic–plasmonic interconnect (HyPPI), on-chip structure, bit flow density, intrinsic modulation, extrinsic modulation.

1. Introduction
The requirement of electronic device down-scaling has led to severe limitations for on-chip communication interconnects [1]–[3]. Likewise, the available bandwidth per compute operation (e.g., Mbps/FLOPS) continues to drop and will likely reach its end at 5 nm technology node according to the 2013 International Technology Roadmap for Semiconductors (ITRS) [4]. The increasing demand for data movement, requires novel interconnect technologies over classical electronic links, particularly in terms of latency, energy efficiency, and integration. These effects are due to the heat generation in current integrated circuits, which can easily melt the chip. Furthermore,
the physical properties of semiconductor material will no longer stay the same once we reach 5 nm feature size.

Photonic interconnects are considered a viable on-chip option since the mid-2000s ITRS roadmap for ICs due to their ability to provide higher bandwidth compared to electrical interconnects. This originates from the parallelism of bosons, which is exploited in Wavelength Division Multiplexing (WDM).

In particular, the low optical attenuation of photonics enables on-chip communications with small power losses. Active optoelectronic devices based on pure photonic elements, however, often rely on optical non-linear physical effects requiring some form of light-matter-interaction (LMI), which is inherently weak, which often leads to significant power consumption. For instance, Batten et al., [5], [6] estimated that about 40% of the total consumption is accounted for tuning the electro-optic modulator, a value that is expected to increase further at lower link utilization rates. Moreover, the sensitive ring-modulator based on high quality factor (Q factor) microrings is photon lifetime limited and bandwidth limited and requires heating to be tuned, which results in challenging dense integration.

The physical reason for the large required footprints and high energy efficiency of active photonic devices stems from the large difference of the dipole momentum between the electronic wave function and a telecommunication photon. This leads to weak LMI, which can fundamentally be enhanced by either increasing the optical density of states, or increasing the field density overlapping with the active material. The latter is deployed, for instance, in the field of plasmonics and metal optics.

A plasmonic interconnect utilizing surface plasmon polaritons (SPPs) allows for diffraction-limited optical modes and high field densities. This enables addressing the footprint challenge of photonics by providing enhanced LMIs that allow the desired functionality (e.g., modulation, signal switching, etc.) to be realized within a small and potentially wavelength-scale device footprint. Nonetheless, plasmonic interconnects are only suitable for short-distance communication purposes since its maximum propagation length is limited to tens of micrometers due to the ohmic loss of the plasmonic coherent electron oscillations at telecommunication frequencies.

In this paper, we propose a Hybrid Photonic Plasmonic Interconnects (hereinafter referred to as “HyPPI”) strategy to combine photonics with plasmonic interconnect technology and show that this combination leads to superior link performance. Furthermore, we provide a comprehensive performance comparison between electrical, plasmonic, photonic and our two proposed HyPPI technologies and compare latency, energy efficiency and throughput performance of our proposed links to pure photonic, plasmonic, and traditional electrical links. We also investigate on-chip crosstalk and bit flow density as on-chip metrics. In particular, we replace the diffraction-limited photonic modulation devices with plasmonic counterparts, but keep the low-loss silicon-on-insulator platform as the passive backbone. In addition, we argue that the light source should be part of the link’s performance consideration and consider it on-chip. This will eliminate the need for an electro-optic modulator (EOM) if it is powerful enough to drive the entire link with high operating frequency. We call this direct light source modulation “intrinsic” (versus “extrinsic” for the EOM use), which allows source power management. Our results show that hybridization of classical photonic and nanophotonic elements is energy efficient (~20 fJ/bit), low latency (< 10 ps for millimeter-range propagation distances), has reasonably long-range transmission (~3 dB loss per length > 10^4 μm) interconnect, and has a bit flow density up to 0.5 Gbps/μm².

2. Link Level Performance

With network-on-chip (NoC) application in mind, the network's performance inherently relies on the characteristics of its building blocks, namely the interconnect links. In this section, we investigate the performances of the HyPPI, plasmonic, and photonic technologies at the link, i.e., point-to-point (P2P), level with respect to latency, energy efficiency, link throughput, and energy-delay trade-off. We also compare with electrical interconnects.
2.1. Point-to-Point Latency

For an electrical interconnect, the P2P latency is defined as the time a single bit of data packet requires to travel from the sender to the receiver. For a photonic or a HyPPI (plasmonic) interconnect, latency includes the time of propagation that a photon (SPP) requires to propagate from the light source across the link to the photoreceiver, which includes the amplification circuit. Thus, for every optical link type investigated here, the latency can be estimated by summing up each device’s individual latency plus the light propagation time.

Given the technologically mature option to use electrical interconnects, the entire link including wires and other components can be combined into a series of RC (\(R = \text{Resistance}, C = \text{Capacitance}\)) lumped elements given by the selected technology node (i.e., 14 nm technology node is considered here based on ITRS) as in (1) and (2), shown below. The on-resistance of the minimum size n-FET \(R_{\text{on}}\), parasitic driver capacitance \(C_{\text{par,dr}}\); load capacitance \(C_L\), optimum number and size of the repeaters \(k_{\text{opt}}\) and \(h_{\text{opt}}\) and the per unit length resistance/capacitance \(r_{\text{int}}/c_{\text{int}}\) in the equations all are borrowed from Rakheja and Kumar [7].

\[
t_{p2p-\text{el}} = 0.69 R_{\text{on}}(C_{\text{par,dr}} + C_L) + 0.69(R_{\text{on}}c_{\text{int}} + r_{\text{int}}C_{\text{fan}})L + 0.38 r_{\text{int}}L^2
\]

\[
E_{\text{el}} = \frac{1}{2}(C_{\text{int}}L + C_o k_{\text{opt}} h_{\text{opt}}) V_{DD}^2.
\]

However, for photonic and HyPPI interconnects, the light source is considered to be on-chip and is kept on at all times during operation for extrinsic modulation. Hence, we can safely neglect the source delay; except for the HyPPI intrinsic modulation because we manipulate the signal directly at the laser not the modulator. Moreover, the latency caused by the source should also be considered for the plasmonic option due to its short propagation length; the entire link structure needs to be repeated using the previous signal output to drive the light source in the next segment in an iterative way. The plasmonic lasers for plasmonic and HyPPI interconnects are borrowed from Lau et al. and Ren-Min et al. [11], [78].

For all four optical link options (photonic, plasmonic, HyPPI-intrinsic and HyPPI-extrinsic) the waveguide delay was estimated based on its material composition, waveguide’s structural shape and propagating light wavelength. Here, we estimate the fundamental mode effective index \(n_{\text{eff}} = 3\) for classic photonic SOI waveguide with moderate loss of 0.5 dB/cm and an Ag-SiO\(_2\) SPP waveguide with \(n_{\text{eff}} = 1.457\) and 440 dB/cm attenuation losses [8], [9]. All the numbers are given for 1550 nm wavelength. Using (3) and (4), shown below, we can calculate the waveguide propagation speed (waveguide latency) and maximum propagation length. For consistency and comparability, similar waveguide cross-sectional shapes were used for the waveguides investigated. Note, the propagation length \(L_{1/2}\) for an SOI waveguide could easily reach few centimeters (\(\sim 9\) cm), which is far beyond the chip-scale. However, for the SPP waveguide, this length is reduced up to 100 microns. Therefore, in the following discussion we assume that photonic and HyPPI interconnects use an SOI waveguide for light transmission and do not need repeaters with the exception of plasmonic links, which repeats every 100 \(\mu\)m.

\[
\text{Propagation speed } v = \frac{c}{n_{\text{eff}}}
\]

\[
\text{Propagation length } L_{1/2} = \frac{\text{total loss}}{\text{unit loss}} = \frac{10 \cdot \lg_{10} \frac{1}{e}}{1}.\]

We studied a wide range of active devices (lasers [10], [11], modulators [21]–[47], and detectors [48]–[72]) with different structures, materials, and technical approaches. The selection for each device is not limited to the device maximum operating speed or estimated energy consumption, but also includes the on-chip footprint, insertion loss, responsivity, cooperation with other devices, and other on-chip characteristics, such as temperature dependence and wavelength spectrum, etc.
For the photonic interconnect, we use a femtojoul level athermal silicon microdisk modulator from Timurdogan et al. [18] and a zero-bias waveguide photodetector from Vivien et al. [58]. For the plasmonic interconnect, we borrowed the modulator and detector from Dionne et al. [26] and Mousavi et al. [64] respectively. For these devices, the latencies are simply regarded as the reciprocal of their operating frequency (60 GHz and 700 GHz, respectively, based on their RC delay).

For the proposed extrinsic and intrinsic HyPPI interconnects, our aim is to reduce the link latency by using ultra-fast plasmonic active devices, while maintaining chip-size long-range propagation via low-loss photonic waveguides similar to those found in silicon-on-insulator (SOI) platforms. In the extrinsic modulation link, instead of using conventional photonic ring-based modulator or the novel microdisk modulator [18], we use an ITO electro-optical modulator using SOI hybrid integration similar to the plasmonic link not only to achieve fast operation and low energy consumption, but also to have an advantage in structural matching since the base structure of the ITO modulator is an SOI waveguide [46]. We should point out that when the link loss is reasonably low, new source technologies can be considered, which offer power savings and unprecedented-fast intrinsic modulation, as found in the emerging nano-cavity plasmon lasers and light-emitting diode (LED) devices [10], [11], [78].

The electrical supporting components (i.e., device drivers) specifications are based on the 22 nm technology node [73]. Since the timeframe of all the optical devices used is in the range of 2009–2014, we chose the technology node level of the year 2012 for predicting electrical components performances. Since the latency of the driver is determined by its RC delay which is proportional to the device capacitance it drives, driver(s) latencies could be calculated using the prediction of Chen et al. (9.5 ps for a 24 fF size modulator) [73].

To obtain scaled down optoelectronic devices, the reduced LMI is compensated by field enhancements and high optical density of states. These wavelength size sources only provide tens to hundreds of micro watts of optical power, and utilize modified rate equations allowing for minuscule power consumption, small footprint, and fast data operation. The latter enables bypassing the external electro-optic modulator altogether. The laser is modulated via an electrical driver (intrinsic modulation). The latencies of the various devices utilized are summarized in Table 1.

Comparing the P2P latency of the various five different interconnects versus link length shows that the plasmonics can indeed outperform pure photonic links due to its higher LMI in the limit of propagation lengths (see Fig. 1(a)). However, the delay for plasmonic links grows when repeaters are required (~100 μm), thus limiting the use of these links to local communication ranges only. The high performance of Silicon photonics can be seen by the relatively short crossover length (~20 μm) with electrical interconnects. To that end, pure plasmonics are a questionable interconnect technology unless other synergies with existing infrastructure such as CMOS would allow us to reduce other technological relevant parameters such as cost. HyPPI can provide much lower P2P latency for both local range due to high LMI and sub-wavelength-scale footprint.
devices and chip-size communications, due to the long-range propagation of silicon photonics ($\alpha_{SOI}$–waveguide $\leq 0.5$ dB/cm).

### 2.2. Energy Efficiency

The link energy efficiency is directly related to the performance of the network and important for both internal and external considerations [4]; internally, it relates to integration density, temperature budgets, and tuning overhead, whereas externally, it relates to overall power consumption and battery life.

While the power consumption of the source in optical links has been included in roadmaps, it has usually been omitted in the literature since it is typically regarded as an off-chip component due to the high power consumption, and subsequent heat dissipation [4], [7], [73]. In this paper, we calculate the required laser power by summarizing the link device losses, photonic-plasmonic coupler loss and the power loss of the entire propagation path to ensure output electrical signal generated by detectors based on responsivity ($R$) still meet the minimum current requirements for the next stage after such losses (see (5), shown below).

In Table 2, we use the same devices when calculating the link latencies and the losses for SOI and we borrow from Li et al. and Berini et al. [8], [9] the values for SPP waveguides. The output minimum current level $I_{\text{min}}$ is assumed to be 50 $\mu$A, which is a moderate value for driving
The bandwidth (operating speed) for photonic link (25 GHz) and plasmonic (60 GHz) are limited by their modulation speed and HyPPI-Intrinsic links (200 GHz) are limited by the switching speed of the light source. Although the light source of the HyPPI-Extrinsic link could be regarded as ON all the time and will not affect the overall operating speed, its speed is also been constrained to 200 GHz for fair comparison. Although some of the performance results are based on optimized theoretical calculation and prediction, exploratory devices have shown to demonstrate over 100 GHz speed and the speed is still increasing [74].

Similar to the latency assumption, the energy consumption for the drivers are also considered and predicted following the same methodology in the previous section. The driver energy consumption is highly based on the applied voltage and the capacitance of device it drives. Therefore, the driver energy equation for a single charge (operation) can be approximated to be \( \frac{1}{2} CV^2 \). Under the same technology node, the voltage levels are usually the same. The driver energy efficiency for HyPPI-extrinsic option shows smaller than the photonic and plasmonic interconnects. This is because the modulator we use has a much smaller device capacitance (size). However, that does not mean the driver of HyPPI consume less energy per second (power) because it also operates at a higher frequency. The energy efficiency, lasing efficiency and power loss in dB of each of the considered interconnects are listed in Tables 2 and 3 broken down based on the same devices used in the latency analysis.

\[
\text{Laser Power per bit} = \frac{l_{\text{min}}}{BW \times \text{Responsivity} \times \text{Efficiency}} \times 10^{\text{loss} \times \frac{1}{10}}. \tag{5}
\]

For comparison, the electrical link is plotted based on its length dependent RC structure mentioned in Section 2.1 [7]. Comparing the energy efficiency with the other four different optical interconnects reveals a significant reduction in power consumption for the links, which use passive SOI waveguide for signal transmission at longer communication distance. However, due to the insertion loss, propagation loss and low optical-electrical conversion efficiency demands that certain amount of energy overhead must be paid which also makes the electrical option more suitable at shorter communication ranges with respect to energy side (see Fig. 1(b)). On the other hand, for a long communication range, the photonic and HyPPI interconnects show much better scaling because of the low attenuation loss of the passive SOI waveguide.

The plasmonic link can provide better performance by sacrificing transmission distance, whereas HyPPI links are able to outperform the photonic and plasmonic links in both metrics. The intrinsic modulation type of HyPPI saves more energy by not using the modulator, but its speed is subject to source modulation speeds. Moreover, the energy efficiency that the model predicts is an upper-bound energy consumption for each link because all the numbers borrowed from other papers as shown in Table 2 are based on the maximum bandwidth of each device.
However, in real applications, the data rate will be dependent on the data rate of the slowest component of the link as it will be the bottleneck. Also, we note that a device operating at a lower data rate will consume less energy.

2.3. Throughput

Throughput of a link, which is commonly defined as a measure of how many units of information (bits) a medium (network) can deliver in a given time period, is another critical metric for evaluating the performance of a link. We can express it as a function of the operating frequency of each device, the propagation length for the signal to be transported from P2P latency and the data packet size \( D \); see (6), shown below and in Fig. 2.

\[
\text{Throughput} = \frac{D}{t_{\text{capacity}} + t_{\text{P2P}}} \tag{6}
\]

\[
\text{Capacity} = 2 \times BW \times \log_2(M) \tag{7}
\]

Fig. 2. Schematic depicting the concept of throughput (which is equal to the data packet size \( D \) divided by the total time to transmit the data through the link and the latter is the sum of the data sending and the data propagation time) and capacity (the maximum message sending rate, which is defined by the Nyquist equation). Moreover, the bandwidth \( (BW) \) in (2) refers to the difference between the upper and lower designed frequencies and is physically bounded by the slowest operating frequency of the devices that the signal passes through. \( M \) in the brackets represents the number of the coding levels and is assumed to be 2 here for the simple case of on-off keying.

![Throughput Diagram](image)

Fig. 3. Comparison of link throughput with (a) a fixed link length (1 mm) and (b) a fixed packet size (64 bits). No WDM was assumed here and all interconnects are using 1550 nm wavelength.

However, in real applications, the data rate will be dependent on the data rate of the slowest component of the link as it will be the bottleneck. Also, we note that a device operating at a lower data rate will consume less energy.

2.3. Throughput

Throughput of a link, which is commonly defined as a measure of how many units of information (bits) a “medium” (network) can deliver in a given time period, is another critical metric for evaluating the performance of a link. We can express it as a function of the operating frequency of each device, the propagation length for the signal to be transported from P2P latency and the data packet size \( D \); see (6), shown below and in Fig. 2.

In reality, \( D \) depends on the different traffic patterns and protocols under a particular computational application. Capacity, as in (7) [bit/s], shown below, refers to the data-carrying ability of a network, indicates that the maximum amount of data that can pass from one point to another per unit time as in (7) (Bandwidth (BW, [Hz])). Fig. 2 shows schematic diagrams for throughput and capacity. Moreover, since the Nyquist equation can only predict the highest capacity for ideal links without noise, we only take half of its results to match with real communication situations. Note, the definition of the term BW is not unique; for instance, Rakheja [7] proposed to define it as the reciprocal of the P2P latency of the entire link. Under this assumption, every bit is sent only after the previous bit has arrived at the end of the link, which results in a rather low capacity, and therefore can be regarded as a lower bound. In contrast, we regard bandwidth of each device in the link as operating as close as possible to its designed frequency, and therefore, our capacity definition establishes an upper bound.

It is worth mentioning that decreasing the data packet size or increasing the link length will both negatively affect the throughput (see Fig. 3(a) and (b)). This can be understood from analyzing (6); decreasing \( D \) for a fixed link length (i.e., P2P latency) will decrease the denominator at a slower rate than the numerator; however, increasing the link length for a given \( D \) also decreases the throughput of the link. Comparing the throughput for the five link types clearly shows an
absolute advantage for HyPPI in providing a high throughput with good response to increasing the data size and high resistivity to increasing the link length. The throughput here is only for single wavelength carrier, and WDM technology will be discussed in future research.

\[
\text{Throughput} = \frac{D}{\frac{D}{\text{Capacity}} + t_{\text{p2p}}} \quad (6)
\]

\[
\text{Capacity} = 2 \times BW \times \log_2(M). \quad (7)
\]

### 2.4. Overall Link Performance

Although the individual link performance we have analyzed thus far can give us an accurate comparison in each aspect, no interconnect option has an absolute advantage in all of the aspects. Therefore, we still need to investigate a more complex figure of merit that includes the energy efficiency; P2P latency and throughput of each interconnect type. Such composite metrics are useful towards defining link merits via exploring tradeoffs among relevant criteria.

In this paper, we use the metric proposed by Martin, Nystroem, and Penzes termed $ET^n$, in which $E$ stands for the energy consumption in the unit of Joules and $T$ for the delay (latency) of the system or network [16]. By multiplying the energy efficiency ($E'$, [J/bit]) by throughput ($Th$, [bit/s]), we get the operational power ($P$ in [W]). Power times P2P latency ($T$, [s]) results in the total energy consumption ($E$, [J]) of the link. Note, $n$ in $ET^n$ metric denotes a weight that represents the degree of correlation between two criteria. For $n = 1$ (see (8), shown below), a decrease in latency is considered as valuable as a reduction in energy consumption at the same proportion, and $n = 2$ is a metric independent of the power supply voltage (see (9), shown below) [75]. For $n = 2$, when comparing the tradeoffs, more energy consumption for small latency improvement in interconnects that a) have a composite performance which is more sensitive to latency, and b) have a relative surplus in its power budget. Note that $n = 0$ is just the energy efficiency metric we already mentioned in the previous section. Therefore, the $ET^n$ metric can be used to optimize the weighted energy consumption and system latency simultaneously, and is consistent with priorities of High Performance Computing (HPC) [76].

\[
\text{Energy Delay Product} = E' \times Th \times T \times T = P \times T \times T = ET \quad (8)
\]

\[
\text{Energy Delay Squared Product} = E' \times Th \times T \times T^2 = P \times T \times T^2 = ET^2. \quad (9)
\]

Interestingly, our results show that the electrical link is superior for link lengths below 5 µm. This is obvious, when considering that the resistance and capacitance is dependent on the link length, and therefore causes smaller latency and energy consumption for short lengths (see Fig. 4(a) and (b)). In contrast for such short lengths, any optical solution has only RC delay from the active components and drivers as a small part, can have fundamental inefficiencies due to a quantum efficiency less than 100%. Moreover, due to the different on-chip device
footprint for each technology option, the curves to the left side (shorter link length) begin at different points. Fig. 4(a) and (b) show absolute advantages for HyPPIs above $10 \mu m$ link length and this again proves that a HyPPI backbone network with local electric interconnects branches is the best alternative.

Additionally, link of about $100 \mu m$ in length can be regarded as a general break-even point for electrical, plasmonic and photonic links. This can be qualitatively understood as the boundary between local to medium range ($< 100 \mu m$) and long range ($> 100 \mu m$) communications and may lead to a design rule for future NoC that are composites of multiple link technologies.

3. On-Chip Performance

3.1. Chip Level Crosstalk

While the P2P latency, energy efficiency, link throughput and the $ET^n$ metrics are important to understand a link’s intrinsic features, crosstalk is a characteristic related more directly to NoC architectures.

Here, we define acceptable distance for crosstalk to be the link length where the energy leakage is 25% from one waveguide to its adjacent neighbor and quantify the crosstalk coupling length of two waveguides. Thus, the total leakage from two adjacent waveguides will cause a maximum 50% energy interference, which is the upper limit of a binary coding system could sustain (below “0.5” as off/zero state and above “0.5” as on/one state). We determine the crosstalk by simulating the eigenmode of the two waveguide structures and solving for the difference between their symmetric and asymmetric modes.

Fig. 5(a) and (b) show that the SOI waveguide mode is cut-off at a dimension of $t_{SOI} = 220$ nm, which can be estimated by the Abbe Limit ($t = \lambda/2n$, where $\lambda$ is the free space wavelength, and $n$ is the effective mode index, which is close to the material index of Silicon at 1550 nm).

In contrast, the plasmonic mode is able to provide sub-diffraction limited (non-cutoff) modes and is hence more suitable for highly integrated chip size ($< 1$ cm) and local-range communication ($< 100 \mu m$). For an SOI waveguide, the maximum coupling length increases rapidly with waveguide dimension and the inter-waveguide spacing (gap) and can easily reach the 1 cm chip size with a gap much smaller than its dimension. This leaves us with two choices; either a longer propagation distance at the cost of a larger waveguide footprint (area) or a denser waveguide arrangement with only local range communication.

Note, the small diameter plasmonic waveguide (e.g., 50 nm radius; see Fig. 5(a)) has essentially the same crosstalk length as to the larger scale plasmonic waveguide (e.g., 300 nm and 500 nm) but is still not practical for on-chip use due to its low optical mode propagation speed amplified by repeaters (i.e., high latency). We deploy an eigenmode solver to obtain the
propagation speed based on the effective mode index for both the SOI photonic and metal embedded dielectric plasmonic waveguide (see Fig. 5(b)).

The propagation speed of plasmonic waveguides reduces significantly when the dimension of the waveguide drops below 100 nm. In addition, based on conclusions from Fig. 1(a), the major component of the P2P latency after increasing the link length is the propagation time through the waveguide. For instance, for a 50 nm radius plasmonic waveguide, the signal only has about half of its maximum speed of light, which will cause an unacceptable delay for the entire link. The latter is another reason why pure plasmonic interconnects might not be a practical option alone despite footprint savings on-chip.

On the contrary, the SOI waveguide has an opposite speed pattern with higher speeds for smaller waveguide dimension (i.e., lower effective index). Although the crosstalk length has little difference between two waveguides of about 300 nm or smaller, the higher propagation speed favors the photonic waveguide resulting in a better link latency and higher throughput.

Furthermore for larger waveguide dimensions, the speed of the photonic waveguide decreases rapidly while the speed of plasmonic waveguide stays almost constant. However, this in return increases the crosstalk length of the photonic waveguide exponentially and can easily reach 1 cm chip size with much less spacing. In other words, even though the propagation time through each single link increases, the gap between two parallel waveguides can be reduced. This would increase the packing density and also would result in higher total throughput density. Note, we only considered the simplest crosstalk model from two adjacent waveguides (25% energy leakage from each waveguide) here for network level performance prediction, which can be regarded as the minimum crosstalk length on-chip since in an integrated waveguide array, the crosstalk will not only comes from the nearest waveguide, but also the second-, third-nearest waveguide(s) etc., which has already been proved by Song et al. [77]. In the future, we plan to include not only multi-waveguides crosstalk but device-to-waveguide and device-to-device crosstalk as well a real optical network structure.

3.2. Bit Flow Density

Crosstalk limits the maximum communication length. On the other hand, the propagation length determines the actual length a signal travels before a repeater is needed. Together these quantifiers define over that distance an interconnect link can deliver a signal before it is undetectable by the receiver and without interfering with other waveguides.

In detail, the dimension (size) of the transmission waveguide and the spacing (gap) between two adjacent waveguides has the most influence on the propagation length and the waveguide-to-waveguide coupling levels respectively. Therefore, it is reasonable to conclude that by increasing the waveguide dimensions and the gap between waveguides, the signal can be delivered a longer distance, however increasing the waveguide dimension and the gap between adjacent waveguides would limits the on-chip footprint (area) and packing density. Thus, we believe there should be an optimal size-to-gap ratio given certain communication requirements. We quantify this footprint-performance trade-off by defining “Bit Flow Density.”

We define Bit Flow Density (BFD) as the number of bits transmitted through a certain chip width (cross-section) to reach a specific required communication length, which is highly related to the size of each device and spacing. The three major elements in chip area calculations we assume are 1) the size of the light sources (laser, LED), 2) modulators (microdisk modulator, PlasMOStor modulator and ITO modulator), and 3) the waveguide area including waveguide-to-waveguide spacing. Note, the area of detectors are not considered here since the output current level at the end of each link is highly dependent on the requirements of the next circuit block and additional amplifiers may be needed as well.

Each of the four different on-chip packing layouts is arranged within a 1 × 1 mm\(^2\) chip area. The laser size (10 \(\mu\)m × 20 \(\mu\)m) for photonic links is taken from the DSENT database [5], and the 20 × 40 × 140 nm\(^3\) core dimension 200 GHz nLED proposed by Lau et al. covered by passivation and silver layers with total on-chip size 80 × 140 × 200 nm\(^3\). The microdisk modulator...
(~5 µm diameter) is simplified to an annulus structure on this 2-D layout, and the space between the ring and the waveguide on one side is much closer (150 nm) than the three other sides for coupling purposes [24]. The larger spacing is used to avoid coupling and its value is determined by eigenmode simulations. The ITO based electro-optical modulator is used in both plasmonic and HyPPI links with a 2 µm device length [10]. The width of the ITO modulator can be regarded as being built “on” the waveguide device and costs no additional flat area due to its compact structure. The detectors used for photonic links and plasmonic/HyPPI links have the on-chip size 10 × 10 µm² and 2 × 2 µm² respectively.

Fig. 6 shows the resulting Bit Flow Density heat map for our four interconnect links. As expected, plasmonic interconnects achieve their highest value in the center of its plot, which indicates that plasmonic links are able to provide high data transmission rates over a certain chip scale with smaller waveguide diameter and gap compared to its photonic counterpart (Fig. 6(a)). However, when the waveguide diameter scale below 50 nm, the crosstalk length and the propagation length respectively will become even smaller than the side length of the PlasMOStor modulator and therefore causes a longer bit flow density.

On the other hand, all the high bit flow density regions are shown on the right hand side (waveguide width > 300 nm) of their plots for both photonic and HyPPI links due to the low crosstalk length (needs more repeaters to propagate enough distance) of SOI waveguides below 300 nm waveguide width. There is only one exception; the lower left corner of the HyPPI intrinsic modulation link is significantly higher than any other regions below a width of 300 nm. This is because the modulators are absent for this interconnect, and therefore smaller waveguide widths and gaps lead to much denser link arrangements.
By comparing all four contour plots, we can conclude that HyPPI (especially intrinsic modulation links) can provide much higher bit flow density rather than both traditional photonic interconnects and plasmonic interconnects. In this regard, intrinsically modulated links combine the best of both worlds; small energy consumption from plasmonic devices and long propagation/crosstalk lengths from photonic waveguides. Such synergies are an example of a reduction in integration and design complexity, which is regarded as a key element to continue scaling on learning curves towards cost reduction and leads to what is previously termed coherent innovation [19].

We note that the BFD and total link energy for chip level photonic links are forecasted to be over 40,000 Gbps/cm², and 10–100 fJ/bit, respectively by the year 2025 [20]. Our results show that to simultaneously achieve the two roadmap goals, HyPPI links are one possible and viable innovative technique.

4. Conclusion

HyPPI hybridizes photonic with plasmonic interconnects to provide a point-to-point link that shows significant improvements in performance relative to that of pure photonic or pure plasmonic links in P2P latency (< 100 ps/cm), energy (∼20 fJ/bit), and combined metrics such as Energy Delay Product and Energy Delay Squared Product. Plasmonic links are limited by their high optical losses whereas; photonic links are bound by high power consumption due to overheads in modulation. Moreover, the sufficient crosstalk length (over 1 cm chip-size) of HyPPI enables dense integration schemes leading to high bit flow density (0.1 ~ 0.5 Gbps/μm², 1~3 orders of magnitude greater than other interconnects at different link length) and higher area efficiency. Such high performance results from technology hybridization appears to be the only technologies at this point in time that can supply both energy and bit-flow-density requirements matching roadmaps for any communication range requirements. These encouraging results will pave the way for designing scalable, low latency, low interference and energy efficient NoCs for future scalable chip multiprocessors.

Acknowledgment

The authors would like to acknowledge the effort of the anonymous reviewers since their timely and constructive reviews improved the paper significantly. They would like to thank the journal editorial board for their timeliness.

References

IEEE Photonics Journal


IEEE Photonics Journal


