SAND2015-4476R

SANDIA REPORT

SAND2015-XXX

Trends in Microfabrication Capabilities & Device Architectures

Todd Bauer
Adam Jones
Anthony Lentine
John Mudrick
Murat Okandan
Arun Rodrigues

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

SAND2015-XXXX

Microfabrication Capabilities & Device Architectures

Todd Bauer
Adam Jones
Anthony Lentine
John Mudrick
Murat Okandan
Arun Rodrigues

Sandia National Laboratories
P.O. Box 5800
Albuquerque, New Mexico 87185

Abstract

The last two decades have seen an explosion in worldwide R&D, enabling fundamentally new capabilities while at the same time changing the international technology landscape. The advent of technologies for continued miniaturization and electronics feature size reduction, and for architectural innovations, will have many technical, economic, and national security implications. It is important to anticipate possible microelectronics development directions and their implications on US national interests. This report forecasts and assesses trends and directions for several potentially disruptive microfabrication capabilities and device architectures that may emerge in the next 5-10 years.

Table of Contents

1. Introduction....6
2. Extreme Ultraviolet (EUV) Lithography....8
3. Nanoimprint Lithography....15
4. Lithography by Directed Self-Assembly....20
5. Channel Engineering....24
6. Memory Integration....28
7. Optical Interconnects....44
8. Neuro-inspired Computing: A New Paradigm....61
9. Conclusions....69

1. Introduction

The last two decades have seen an explosion in worldwide R&D, enabling fundamentally new capabilities while at the same time changing the international technology landscape. For example, in microelectronics, consistent with the predictions of Moore's Law, the number of transistors on an integrated circuit has doubled roughly every 18 to 24 months. The advent of technologies for miniaturization and electronics feature size reduction, heterogeneous integration, and advanced processing/fabrication techniques have thrown the door wide open with respect to innovative devices that may be fieldable in the next 5-10 years. These and other emerging developments will have many technical, economic, and national security implications. It is important anticipate possible microelectronics development directions and their implications on US national interests. With this in mind, this report forecasts and assesses trends and directions for several potentially disruptive microfabrication capabilities and device architectures that may emerge in the next 5-10 years. This report is informed by the current state of the art for integrated circuit architectures and fabrication but this report is not intended to cover the convential, which has been done elsewhere. For an excellent review, see *The Future of Computing Performance* by the National Research Council of the National Academies.[1]

In this report, we evaluate a sampling of emerging process technologies and device architectures, with inclusion based on our determination of which technologies may have dramatic and surprising impact in the next five to ten years and which are areas of local expertise to maximize our value added. Among process technologies, we cover the current state of the art and future research directions for extreme ultraviolet (EUV) photolithography, nanoimprint lithography (NIL), lithography by directed self assembly (DSA), and channel engineering. EUV lithography may seem like an odd inclusion since it has been in development for over a decade and it is exceptionally expensive to execute research in this area. But, light sources have recently improved and EUV lithography is nearly ready for high volume manufacturing. Once it is being used in production, we anticipate that EUV lithography will enable non-linear scaling advantages when paired with techniques like double and triple patterning. Innovations related to EUV are not likely to occur in a university laboratory, but they may occur in factories under strong government influence. NIL and DSA lithography are intriguing because they are relatively immature technologies that hold great promise for their potential to pattern small features at relatively low cost. Likewise, channel engineering is an area where creative research executed in a university lab could have profound impact on transistor performance. Among architectures, we focus on new approaches to memory integration, optical interconnects that use photons rather than electrons to transmit information, and neuro-inspired computing, which seeks to mimic the function of the human brain

---

[1] The National Resource Council of the National Academies, The Future of Computing Performance, S.H. Fuller and L.I. Millet, Eds. Washington D.C., 2011.

and the nervous system to process information. New approaches to memory integration and optical interconnects are ready to break out as widely deployed technologies. New memory integration approaches will impact everything from home computing to high performance scientific computing. It may also enable other new architecture approaches like neuro-inspired computing. By improving data transmission speed and reducing power, optical interconnects will impact data centers, enterprise computing and cloud computing, and high performance computing. Neuro-inspired computing is compelling because it may revolutionize how we analyze big data and how automated control systems work. Neuro-inspired computing need not be a stand-alone reinvention of computing. It may ultimately be integrated as a complement to conventional sequential approaches rather than as a replacement. In this report, we leverage the recent Neuro-Inspired Computational Elements (NICE) Workshop to characterize the current state of the art and identify future trends.

In terms of tech surprise, it is possible that the the least-expected high-impact advances occur at the intersection of new process technologies and new architectures. We encourage the reader to keep that in mind while considering their own analysis and investments.

Lastly, our experience as engineers and scientists that have served both government and industry is that innovation and tech surprise will continue to be driven by creative, highly-trained people who are given a directive to make something work with the tools at hand. With that in mind, analysts should consider the human talent required to execute cutting edge technical work, and recognize that a critical mass of highly-educated, well-resourced people will be the result of coordinated government action to establish universities and research centers.

2. Extreme Ultraviolet (EUV) Lithography

Overview

Several next generation lithography (NGL) technologies are under development in order to drive transistor sizes to even smaller dimensions. Extreme ultraviolet (EUV) lithography is among the more mature of these NGL techniques and chip manufacturing industry leaders including Intel, TSMC, and ASML have invested significant resources towards its development. EUV is the logical next step for reducing the pitch of patterned structures, and, in contrast to other NGL techniques, strategies for managing defects for this optical projection lithography approach are well understood.

Current optical lithography processes used in production-scale semiconductor microfabrication employ a deep ultraviolet (DUV) illumination source for pattern development and transfer. The minimum achievable feature size or resolution, R, is related to the illumination source wavelength, λ, according to:

$$R = \frac{k_1 \lambda}{NA}, \tag{1}$$

where k_1 is a process-dependent constant and the numerical aperture NA describes the orientation and composition of the optical beam focusing system. State of the art production-scale processes have used a λ = 193 nm DUV light source for nearly two decades. Several process-related innovations have decreased k_1 and increased NA in equation (1) to drive production to the 14 nm feature node with excimer-based DUV light sources, but further reductions in feature size and pattern pitch are a constant challenge. Conceptually, EUV lithography is a straightforward approach where R is reduced by using a shorter wavelength light source: 5 nm < λ < 15 nm. For conservative production parameters of $k_1 = NA = 0.3$, a λ = 13.5 nm EUV illumination source would theoretically be able to produce feature sizes at the present state of the art node, R = 14 nm. However, several challenges exist and need to be overcome in order for EUV to be implemented in production-scale manufacturing.

Current Challenges and State of the Art

Before describing the key challenges to EUV adoption, it will be helpful to describe the system operation in more detail. An illustrative schematic is shown in Figure 1[2]. The most common EUV generation sources for lithography are laser-produced plasma (LPP) systems. LPP systems consist of a CO_2 laser that excites tin droplets to generate a plasma, which in turn emits EUV photons. These sources are incoherent, where photons are emitted from a point source in all directions. Careful optical design is therefore required to create a focused intermediate beam to be delivered to the target wafer. This optical system primarily consists of a collection lens, which focuses EUV photons on an

[2] C. Wagner and N. Harned, "EUV lithography: Lithography gets extreme," *Nat. Photon.*, vol. 4, pp. 24-26, Jan. 2010.

intermediate spot, and a series of reflective mirrors that function to steer the EUV beam to the target wafer.

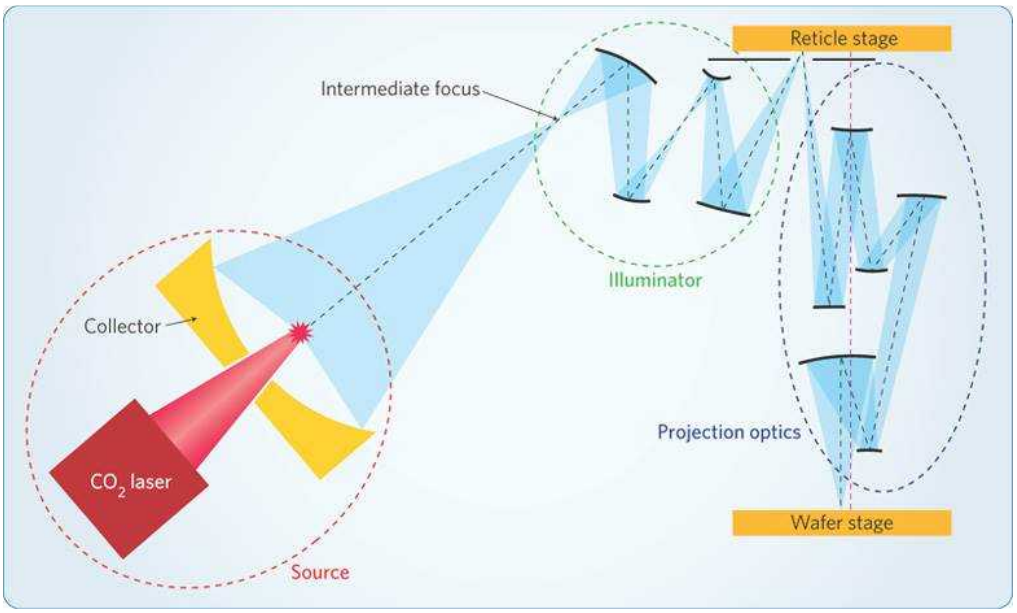

Figure 1: Schematic of an EUV lithography system comprised of EUV source, reticle/mask, and focusing optics[3].

In addition to EUV photons, the generation process also creates other byproducts which contaminate the collection lens. Therefore the collection lens must be continuously cleaned to maintain high reflectivity and collection efficiency. EUV leader ASML has demonstrated an *in situ* collector cleaning process resulting in collector mirror lifetimes approaching one year, satisfying reliability concerns for this element[4]. Because nearly all materials absorb EUV energy, two more restrictions are placed on an EUV lithography system: (1) reflection-based focusing optics are required because no compatible EUV-transmitting materials have been developed, and (2) a vacuum environment free of contaminants is needed. At present, the best-performing reflective alignment mirrors are relatively complex Bragg reflectors each consisting of ~40-50 Mo/Si thin film bilayers patterned with absorbing mask structures. These mirrors have reflection coefficients approaching 0.7. The focusing optical system contains up to six of these mirrors within the beam path in order to separate incident and reflected light rays, yielding a theoretical maximum collection efficiency of $(0.7)^6 = 12\%$.

While minimum feature sizes achieved with EUV patterning have quickly reached those achieved with DUV-based processes, the relatively low flux of EUV photons delivered to resists at the wafer level are not yet high enough for production-scale implementation. Low EUV generation efficiency, resist sensitivity, and the aforementioned collection

---

[3] Ibid.

[4] R. Peeters, "EUV lithography: NXE platform performance overview," *Proc. SPIE*, vol. 8679, doi:10.1117/12.2010932, 2014.

efficiency all contribute to this shortfall and are the main challenges to EUV implementation in production-scale lithography processing. Strategies for overcoming these limitations will be presented in the following section. The most recent state of the art EUV lithography system reported exposing 1,022 wafers to sustained 90 W EUV power over one 24 hour period[5]. Japan-based manufacturer Gigaphoton Inc. has announced a 140 W source operating at 50% duty cycle, though implementation of this source into wafer processing operations has yet to be demonstrated[6]. High resolution patterning has yet to be demonstrated on such high power, high throughput EUV systems.

Research Directions

Progress on source brightness, photoresists, and optical path components will all impact the success of EUV lithography. Mask infrastucture to instantiate pellicles, inspect, and clean EUV photomasks also needs to be developed. We cover those topics in this section.

Despite improvements in the last few years, the most significant challenge to EUV lithography continues to be the brightness of the EUV source. For the current LPP systems, gradual improvements have come from optimization of the operating parameters of laser/Sn droplet delivery. Gigaphoton Inc. implemented a second, lower energy pre-pulse laser to transform the droplet into a fine mist of micrometer-sized Sn fragments which absorb and convert subsequent CO_2 laser radiation more efficiently. This is depicted schematically in Figure 2(a). Separate optimization of the initial Sn droplet size and pulse laser energy yielded three- to ten-fold increases in EUV generation efficiency[7]. Industry leader ASML similarly emphasizes the importance of these parameters for obtaining high EUV photon flux. Detailed physical models and numerical simulations give a more accurate description of pre-pulse laser energy absorption and Sn target vaporization; these studies conclude that further improvements in EUV generation efficiency are attainable[8].

An intriguing approach to improving EUV power is to replace the LPP source with a free electron laser (FEL). FELs have been under development since the 1970s and regularly generate kilowatts of EUV power, two orders of magnitude more than levels aggressively targeted by LPP sources. FEL systems consist of a relatively large electron accelerator which sends high energy electrons through an undulator tuned to generate photons of a specified energy, as shown in Figure 2(b). These units have large space requirements (hundreds of square meters) and are highly cost-intensive: a 2012 concept study by Helmoltz Zentrum Berlin and Zeiss estimated the cost of a manufacturing-scale FEL

---

[5] http://semiengineering.com/manufacturing-bits-feb-24/

[6] http://electroiq.com/blog/2015/02/gigaphoton-achieves-continuous-140w-euv-light-source-output-at-50-duty-cycle/

[7] H. Mizoguchi, *et al.*, "Sub-hundred Watt operation demonstration of HVM LPP-EUV source," *SPIE Proc.*, 2014.

[8] T. Sizyuk and A. Hassanein, "The role of plasma evolution and photon transport in optimizing future advanced lithography sources," *J. Appl. Phys.*, 2013.

system at over $200 million[9]. The promise of high EUV photon flux and the potential for higher energy (smaller λ) radiation have drawn new focus to this application of FEL technology, and two sessions were devoted to this topic at the 2014 International Workshop on EUV and Soft X-Ray Sources. This level of interest suggests FEL could become an important EUV-enabling technology in the coming decade, if the cost can be brought down.

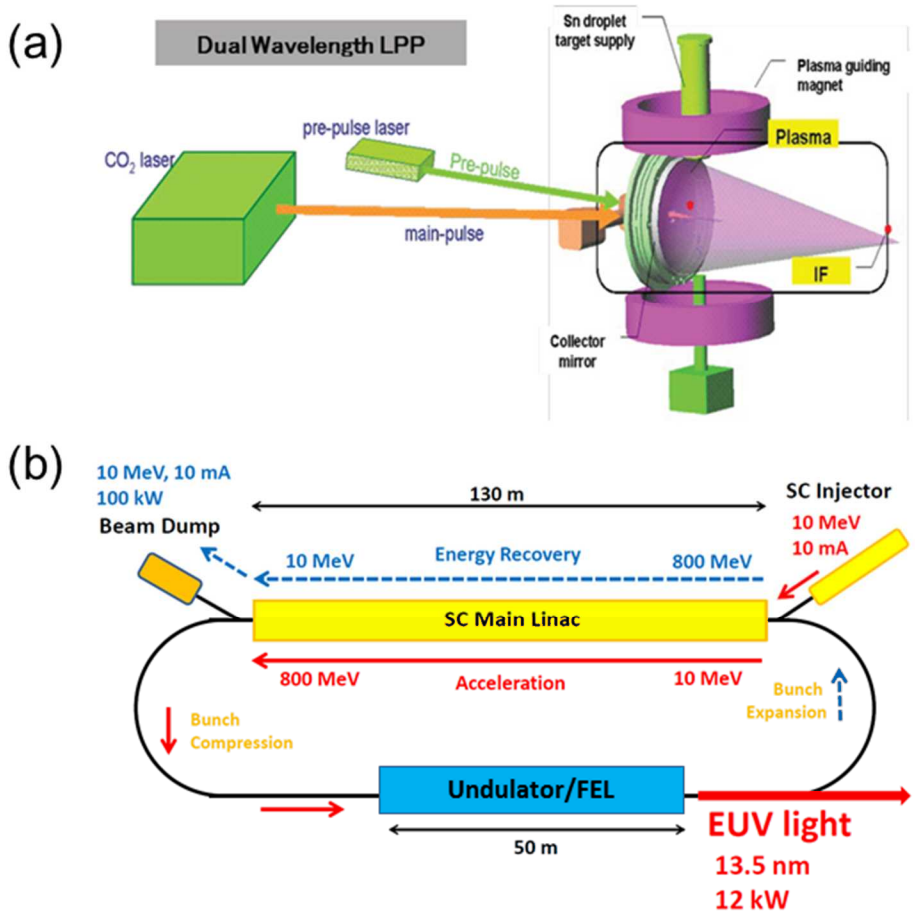

Figure 2: (a) LPP generation scheme including a pre-pulse laser[10] (b) FEL generation system with large physical dimensions and high EUV power[11].

---

[9] D. Turke, *et al.*, "Concept study on an accelerator based source for 6.x nm lithography," *2012 International Workshop on EUVL.*

[10] H. Mizoguchi, *et al.*, "Sub-hundred Watt operation demonstration of HVM LPP-EUV source," *SPIE Proc.*, 2014.

[11] E. Kako, *et al.*, "Development of Superconducting Accelerator with ERL for EUV-FEL," *2014 International Workshop on EUV and Soft X-Ray Sources*.

Tuning the properties of the EUV-absorbing resist can reduce the required source power, but a complex relationship exists between the resist sensitivity, line edge roughness (LER) of the final pattern, and image resolution. One approach to increasing sensitivity, corresponding to a lower required EUV flux during exposure, is dissolving metal oxide nanoparticles inside the polymer resist film. In one such instance, resist films with core-shell nanoparticles consisting of ZrO_2 and HfO_2 cores and dimethylacrylate shell ligands achieved sensitivities lower than 3 mJ/cm^2, well below the industry target of 10 mJ/cm^2. However the resulting 20 and 30 nm-wide line-space patterns had LER between 5 and 7 nm, well above threshold values needed for sub-14 nm feature sizes[12]. A more systematic and fundamental approach consists of modifying the local chemistry of the resist components by treating the functional units as modular pieces to be "plugged in" to different locations within the polymer structure, as in Figure 3(a). Such an investigation by Dow Electronic Materials did not find a champion resist but did identify key relationships between the polymer backbone/leaving group combination and resist performance characteristics[13]. Shiratani *et al.* employ a similar approach, relating glass transition temperatures and quencher doses of a series of resists to their performance metrics[14]. They ultimately achieved 4 nm LER for 18 nm half-pitch line patterns, but the resists still suffer from the LER/sensitivity tradeoff with sensitivities > 40 mJ/cm^2. Achieving a resist capable of patterning low LER, sub-10 nm features and having sensitivity < 10 mJ/cm^2 is an ongoing challenge in this field.

Regarding EUV optical components, the amount of EUV power arriving at the wafer is reduced by non-unity reflection coefficients of the focusing mirrors present in the system. As shown in Figure 3(b), theoretical calculations suggest that nanometer-scale intermixing occurring at each Mo/Si interface reduces reflectivity by up to 20% relative to the hypothetical case with no mixing[15]. Strategies for avoiding this nanoscale intermixing have not been thoroughly investigated, but alleviating a portion (or all) of this phenomenon would directly increase the amount of delivered EUV power. An interesting approach to increasing reflectivity is to introduce nm-scale voids into the multilayer structure. Lee *et al.* show that introducing spherical voids into the Si layers would increase mirror reflectivity from 73% to 83%[16]. For a six-mirror system, such an increase corresponds to a total reflectivity of $(0.83)^6 = 0.33$, a three-fold improvement over a system of mirrors each having 70% reflectivity (0.12).

---

[12] S. Chakrabarty, *et al.*, "Increasing sensitivity of oxide nanoparticle photoresists," *SPIE EUV 2014.*

[13] O. Ongayi, *et al.*, "Effect of leaving group design on EUV lithography performance," *SPIE EUV 2013.*

[14] M. Shiratani, *et al.*, "Novel EUV resist materials for 16nm half pitch and EUV resist defects," *SPIE EUV 2014.*

[15] V. Philipsen, *et al.*, "Actinic characterization and modeling of the EUV mask stack," *SPIE Proc.*, 2013.

[16] Y.-M. Lee, *et al.*, "Void-based photonic crystal mirror with high reflectivity and low dissipation for extreme-ultraviolet radiation," *J. Micro/Nanolith. MEMS MOEMS 2013.*

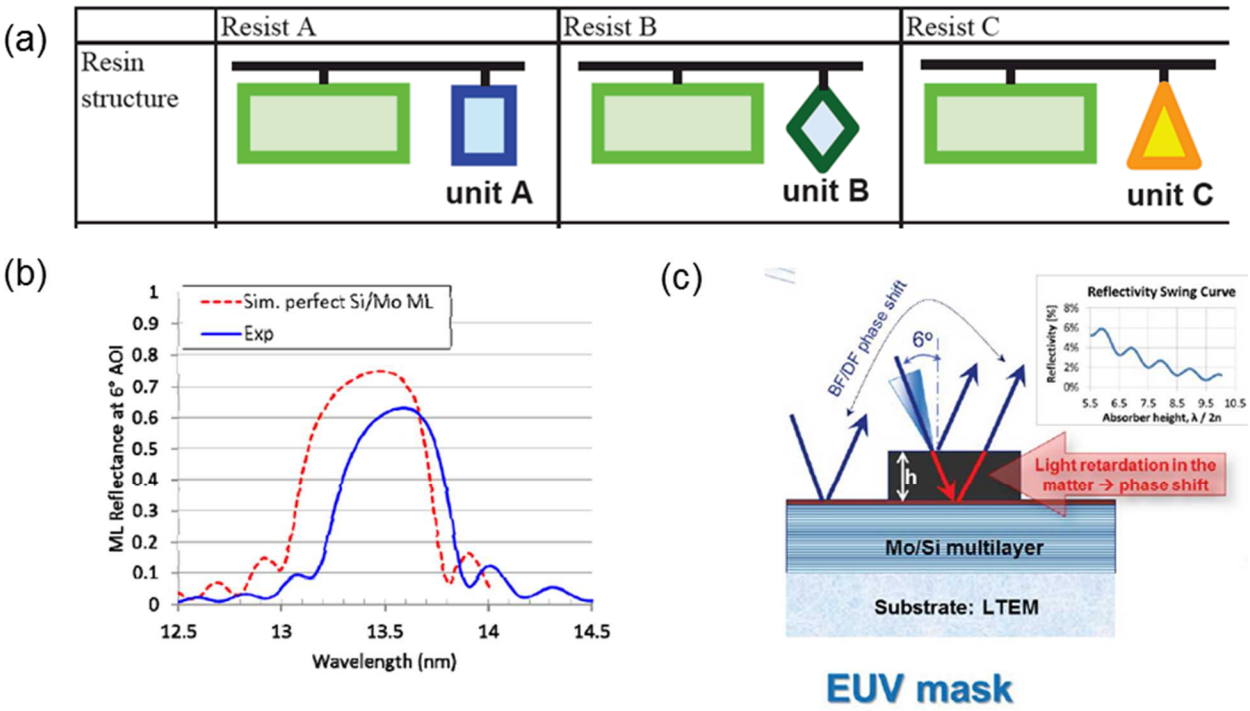

Figure 3: (a) Modular approach to resist formulation with different chemical constituents at on the polymer backbone[17]. (b) Significant improvement in mirror reflectance without Si/Mo intermixing [18]. (c) Focused cartoon of the full mirror-plus-mask assembly, with mask height *h* for optimization[19].

Recalling equation (1), increasing the numerical aperture *NA* is another strategy for achieving higher resolution. This approach places additional constraints on the optical focusing system and requires more precise control over the absorber mask layers used for patterning. Tantalum-based thin films such as TaN and TaBN have been used historically as absorbing structures, and their thicknesses must be above ~50 nm in order for complete EUV absorption. At this length scale, diffraction of EUV light within the multilayer reflector structure results in shadowing of the outgoing EUV rays by the periodic absorber stacks. Efforts to reduce the absorber layer thickness have included optimizing the Ta-based absorber layer thickness[20] [Figure 3(c)] and substituting Ni as the absorber material due to its stronger EUV absorption and corresponding thinner

---

[17] Ibid.

[18] V. Philipsen, *et al.*, "Actinic characterization and modeling of the EUV mask stack," *SPIE Proc.*, 2013.

[19] N. Davydova, *et al.*, "Experimental approach to EUV imaging enhancement by mask absorber height optimization," *SPIE Proc., 2013.*

[20] N. Davydova, *et al.*, "Experimental approach to EUV imaging enhancement by mask absorber height optimization," *SPIE Proc., 2013.*

layers[21]. Following up on these approaches should further improve EUV collection efficiency and patterned image fidelity.

Continued development of the mask production and qualification processes will also be necessary before EUV is fully integrated into wafer production. Efficient production of the total mask assembly (blank substrate, reflector multilayers, protective layer, and absorber mask) with sufficiently low defect levels has yet to be realized. Furthermore, mask inspection tools for qualification of these components are not production-ready. Current tools are either too slow or stretched to the ends of their operating limits, and desired actinic-based inspection methods are still in the development phase. These infrastructural challenges will need to be solved for large scale EUV implementation.

While the main efforts in EUV lithography development have been focused on EUV generation, wafer level power, and focusing optics, other challenges to production-scale EUV lithography exist and solutions are under development. A sufficient pellicle, the thin film which keeps contaminants away from the imaging focal plane, is still a work in progress. EUV transparency goals have been met, but significant improvements in withstanding EUV damage are needed to improve the pellicle lifetime. Mask cleaning procedures also need to be refined in order to avoid degrading or otherwise affecting the mask and other components.

Finally EUV lithography is unique among the topics in this report in that research towards its development and deployment is inherently and inescapably expensive and progress will continue to be the domain of large corporations or other entities with large financial resources.

---

[21] A. Rastegar, *et al.*, "Study of alternative capping and absorber layers for extreme ultraviolet (EUV) masks for sub-16nm half-pitch nodes," *SPIE EUV 2014.*

3. Nanoimprint Lithography

Overview

The inaugural embodiment of nanoimprint lithography (NIL) shared all but feature size with the age-old process of hot embossing[22]. In its most basic form, a rigid template whose surface has been modified to contain the features to be transferred is mechanically pressed against a substrate bearing a polymeric resist layer. Raising the temperature above the glass transition (T_g) of the resist results in conformal deformation of the polymer to the surface relief pattern of the rigid template. The template and substrate are then cooled below the T_g of the polymer followed by physical separation of the joined entities leaving the inverse of the template features in the now-rigid polymer resist; a schematic of the process is shown in Figure 3a. The seminal work by Chou demonstrated sub-10nm resolution from the onset with transfer of 5nm features later[23] thus providing a proof-of-concept of the technology's resolution capabilities and its potential as an approach for Next Generation Lithography (NGL). Over the years, several variations of NIL have arisen targeting numerous application spaces with reduction of defectivity and increase in throughput as primary foci. Importantly, the resources required to demonstrate a state of the art patterning process with NIL are relatively modest and in fact many university labs are equipped and staffed to do important NIL work.

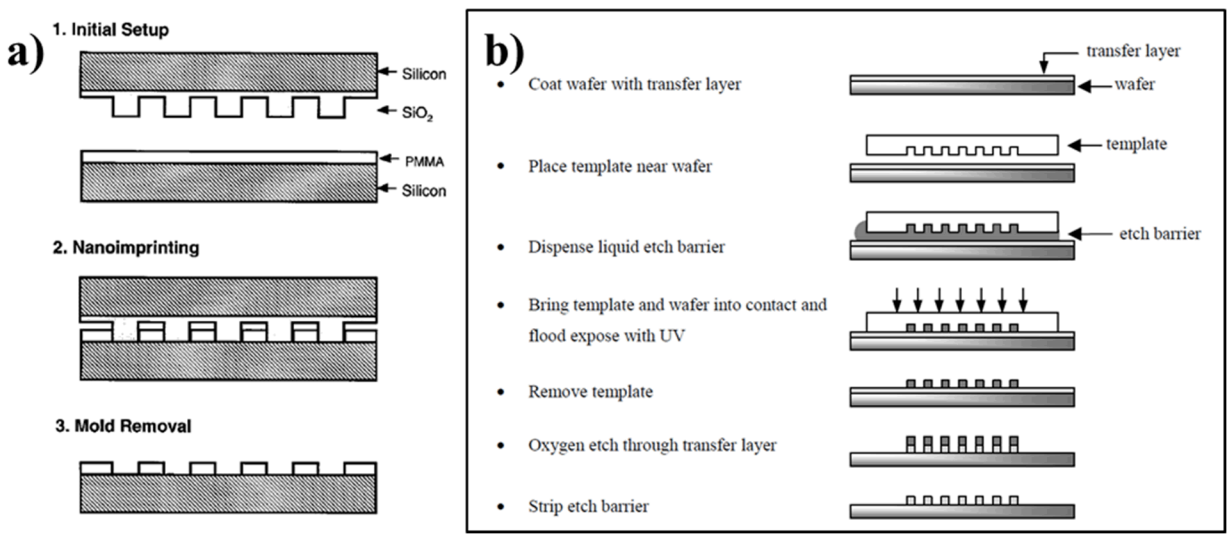

Figure 3: Schematic representations of a) thermal NIL and b) S-FIL.

---

[22] S.Y. Chou, *et al.*, "Imprint of sub-25 nm vias and trenches in polymers," *Appl. Phys. Lett.*, 1995.

[23] M. Schvartzman and S.J. Wind, "Robust pattern transfer of nanoimprinted features for sub-5 nm fabrication," *Nano. Lett.*, 2009.

In UV NIL, a UV-curable liquid resist is deposited on the wafer followed by mechanical pressing of a UV transparent template onto the substrate. A UV source is then used to expose the structure thereby crosslinking the resist in a process not unlike contact lithography with a negative photoresist. The primary advantages of UV NIL over thermal counterparts are throughput and overlay. In the thermal approach, the resist layer in the region of the template must be uniformly heated to a temperature typically on the order of 150°C prior to compression and subsequently cooled prior to release. This heating process bounds the minimum time scale on which the lithography process can occur as uniformity is directly proportional to thermal mass. The time scale of the curing process in UV NIL is rather similar to the exposure times in standard photolithography processes where wafer handling and alignment comprise the bulk of the processing time. Improvement in overlay is observed as coefficient of thermal expansion (CTE) mismatch between the resist, substrate, and template manifest as magnification errors that degrade overlay accuracy; again, UV NIL's similarity to photolithography yields improved overlay accuracy. UV NIL further advanced with development of step and flash imprint lithography (S-FIL)[24]. In S-FIL (Figure 3b), a template whose size is comparable to the field size of an optical stepper is used with an appropriate amount of resist locally dispensed just prior to compression. This enables a repeatable process capable of leveraging the technologies utilized in today's most advanced lithography tools and moves NIL one step closer towards viability as an NGL candidate.

Master/daughter schemes utilize a rigid master template (e.g. quartz or silicon) from which a daughter template (typically PDMS) is fabricated via a thermal imprint process[25]. The PDMS template is then used to imprint the design features into resist coated substrates reducing wear on the expensive master templates. The use of daughter templates is unique in that it enables processes that typically produce prohibitive template degradation. For example, in microcontact printing (μCP)[26], the template is coated with a transferable ink and then pressed against a clean substrate to transfer the design pattern. Over a significantly shorter lifetime than expected with thermal or UV approaches, the template experiences significant fouling and requires replacement. Microtransfer molding (μTM)[27] also relies on coating of the template, but here, the resist is spun onto the template followed by mechanical pressing of the template and clean substrate and, finally, curing of the resist material prior to release. Both of these methods rely on coating of the template and, thus, significantly reduce template lifetime.

---

[24] D.J. Resnick, *et al.*, "Step and flash imprint lithography," *Mater. Today*, 2005.

[25] C. Peng, *et al.*, "High fidelity fabrication of microlens arrays by Nanoimprint using conformal mold duplication and low-pressure liquid material curing," *J. Vac. Sci. Technol. B*, 2007.

[26] G.P. Lopez, *et al.*, "Patterning and imaging of two-dimensional patterns of proteins adsorbed on self-assembled monolayers by scanning electron microscopy," *J. Am. Chem. Soc.*, 1993.

[27] X.-M. Zhao, *et al.*, "Fabrication of 3D microstructures: Microtransfer Molding," *Adv. Mater.*, 1996.

Technology Challenges

The primary barriers to adoption of NIL technologies include defectivity, overlay accuracy, throughput, thickness of the residual resist layer, and the use of 1X templates. Defectivity in NIL arises from several sources including standard particle defects on the template and wafer surfaces in addition to non-fill defects where features in the template fail to fill and plug defects where imprinted features fail to release thereby 'plugging' a recess in the template; an error that transfers to all subsequently printed areas. Defectivity levels as low as $0.3/cm^2$ were reported by Toshiba[28] with a more detailed study by Higashiki reporting $10/cm^2$ [29]. While the former puts NIL's numbers in line with current immersion lithography tools, the reproducibility and sustainability of defectivity levels capable of meeting 22nm node requirements remains to be seen.

Overlay errors arise primarily as magnification errors caused by shrinkage of the resist layer during the curing process and/or CTE mismatch, depending on the process used, while the thickness of the residual resist layer following imprint depends heavily on process. The use of 1X templates leads to a significant increase in template fabrication cost as it must be fabricated using the very methods to which NIL is an alternative. Throughput is currently low when compared to the most advanced photolithography tooling, but generally appears to be an economically driven limitation.

Current State-of-the-Art

Jet and Flash Imprint Lithography represents the current state-of-the-art in UV NIL for large-scale production. Here, an inkjet head is used to disperse small droplets (on the order of picoLiters) across the image field of the mask with the local volume of liquid dispensed varied based on local feature density. The result is a significant reduction in process time, residual layer thickness, required mechanical pressure, and waste as fluid volume and flow are minimized. The report by Higashiki vetted NIL for the 22nm node with defectivity levels as low as $10/cm^2$, fill times under two seconds, 2nm line edge roughness, and 10nm overlay accuracy.

Recently, the combination of J-FIL and directed self-assembly (DSA) has emerged as a route forward in producing low defectivity, high resolution printing. In this scheme, a thin layer of resist is imprinted in the usual way. A block copolymer is then deposited on the substrate whose components are differentially attracted to the resist layer. This causes the block copolymer to self-assemble in a configuration determined by the lengths of the constituent polymer chains[30]. Hitachi Global Storage (HGST)[31] recently used this

---

[28] T. Higashiki, *et al.*, "Nanoimprint lithography and future patterning for semiconductor devices," *J. Micro/Nanolith. MEMS MOEMS*, 2011.

[29] M. LaPedus, "Toshiba claims to 'validate' nano-imprint litho," *EE Times,* October 2007.

[30] X. Man, *et al.*, "Organization of block copolymers using nanoimprint lithography: comparison of theory and experiment," *Macromolecules*, 2011.

approach to create isolated islands of magnetic media of 10nm diameter over large areas potentially enabling the production of hard disk drives with double the bit density of today's hard drives.

Of additional interest outside of the semiconductor industry is the potential for complementary lithography. The concept allows for simultaneous patterning of large and small features in a single lithography step using transparent templates partially coated with chrome. Use of UV curing would then allow photolithographic definition of large features with fine features left to the imprint process. This further enables process optimization as parameters could be independently modified based on local densities and feature sizes.

Finally, the ability to imprint multi-level structures in the resist layer during a single lithography step represents a significant advantage for NIL. Direct imprinting of 3D structures in a polymeric resist layer is possible following fabrication of a suitable template with applications ranging from the production of efficient diffractive optical elements (DOEs) to streamlined dual damascene processes.

Future Directions

Nanoimprint lithography represents a method whereby next generation feature sizes may be printed in a cost-effective manner. While defectivity will likely remain prohibitive to adoption of NIL as the de facto lithographic technique in future CMOS nodes, the technology displays a winning value proposition for defect-tolerant applications, or for applications where low yields are tolerable. The aforementioned work by HGST demonstrated the suitability of NIL to the cost-sensitive storage market while a collaboration between Canon and Toshiba seeks to develop a NAND memory fabrication process at the 15nm node using NIL[32]. Development of optical applications is supported by the recent statement by Schott AG who envisions nanostructuring as a 'technology platform for the future development of innovative applications and products"[33].

The ability to easily imprint 3D structures in a single lithographic step makes NIL disruptive in some application spaces including dual damascene processes and the fabrication of DOEs. Additional applications to which NIL is uniquely suited include the fabrication of lenslet arrays, formation of superhydrophic surfaces, surface modification of LEDs to enhance directivity, surface modification of thin film solar cells to enhance efficiency, formation of anti-bacterial and anti-fouling surfaces, and fabrication of moth-eye antireflective coatings, to name a few.

---

[31] T.R. Albrecht, *et al.*, "Bit patterned media at 1 Tdot/in^2 and beyond," *IEEE Trans. On Magnetics*, 2013.

[32] P. Clarke, "Report: Toshiba eyes imprint litho for 15nm Flash," *Electronics360*, March 2014.

[33] "Next-generation lithography: Schott shooting to be a player in nanoimprint lithography," *LaserFocusWorld,* January 2013.

Compatibility with resolution enhancement techniques such as directed self-assembly (DSA), spacer defined double patterning (SDDP), and sequential infiltration synthesis (SIS) will enable NIL to keep pace with competing lithographic techniques while the low cost of ownership, ability to print in 3D, and the potential for patterning of functionalized polymeric materials will support widespread adoption of the technology in select market segments.

4. Lithography by Directed Self-Assembly

Overview

In this section we discuss directed self assembly, or DSA, which is complementary to optical lithography because it requires a template or pre-pattern for block copolymers to align to. Starting with templates formed by advanced optical lithography, DSA has been demonstrated to pattern features as small as the 15nm range[34], which puts it close to EUV lithography. However, some key challenges remain before it can be used for production. Here we introduce DSA at a high level and then discuss the technical challenges of implementing it and advancing the capability. As a complementary technique, DSA relies on a template formed by other process and we do reference DSA on the section on nanoimprint lithography.

The fundamental element of the DSA approach is the block copolymer[35]. At the highest level, a block coplymer is comprised of two chemically dissimilar polymer chains that covalently link. The strong covalent bond leads to excellent miscibility in solvent, and upon heating or other input of energy the diblock copolymer forms into arrays of nanostuctures called microdomains. The solvent evaporates leaving just the microdomains. Subsequent processing like plasma RIE or UV exposure can be used to volatilize one component of the block coploymer, leaving behind the other component to serve as a soft mask for pattern transfer. Constituent polymers include materials like poly(methyl methacrylate) (PMMA) and polystyrene (PS), which in fact form a block copolymer that is a leading candidate for DSA for pattern transfer.

Directed self-assembly at the nanoscale is dependent on the characteristics of block coplymer and the assembly conditions. Many block copolymers can assemble to form large, three dimensional multiblock structures and it is these bulk structures that have been studied historically[36]. However, for lithography and pattern transfer purposes, we seek to form flat, thin film-like structures with lateral ordering leading to a desired pattern formed in two dimensions. Any DSA approach for pattern transfer must be easy to implement in the typical microfabrication environment and the resulting pattern must be adequately robust to microfabrication processes. To that end, block copolymer can be processed like optical photoresist (spin-on application, hot plate bake, edge bead removal, etc.) and the block copolymer chemistries are highly tunable to a specific application. For instance, the copolymer can be synthesized to have affinity for specific surfaces and for specific arrangements on a given surface. The copolymer can be designed to arrange vertically, dependent on the strong affinity of a specific microdomain for a specific

---

[34] B. Rathsack *et al.*, "Pattern Scaling with Directed Self Assembly Through Lithography and Etch Process Integration", *Proc. SPIE*, vol. 8323, doi: 10.1117/12.916311, 2012.

[35] C.J. Hawker and T.P. Russell, "Block Copolymer Lithography: Merging "Bottom- Up" with "Top-Down" Processes," *MRS Bulletin*, vol. 30, pp. 952-966, December 2005.

[36] Ibid.

surface, and in the process constrain the orientation of other microdomains, leading to the preferential assembly of shapes like linear chains or cylinders.

DSA for pattern transfer uses the carefully tuned surface energies of block copolymers in combination with templated surfaces to form a complementary surface pattern that can subsequently be transferred into another film or the substrate. Templates are formed as physical structures (line, tenches, holes, or posts) that rely on mechanical confinement to array the block copolymers (graphoepitaxy), or by surface modification using chemistry to change the chemical affinity for coblock polymers (chemoepitaxy). Either end of the block copolymer is targetted to have higher affinity for the physical or chemical template, and ordered structures are formed as a result of self assembly according to the characteristics of the template. The small size of the block copolymer combined with subsequent selective microfabrication processes (to remove one component of the block copolymer, or the template and substrate without removing the arrayed polymer) allow the formation of structures that are smaller in feature size than is possible with the process used to form the template (typically optical lithography or nanoimprint lithography). An example of a DSA process flow by chemoepitaxy is shown in Figure 4[37].

Technology Challenges and Future Directions

Key challenges for DSA include shrinking the pitch of patterned structures and forming shapes necessary for integrated circuits[38], and reducing defectivity.[39]

The upside of DSA for IC fabrication is shrinking the pitch of patterned structures without requiring EUV lithography but there are limits to what is currently possible. As previously mentioned, a commonly used block co-polymer is poly(styrene-block-methyl methacrylate) (PS-b-PMMA). PS and PMMA monomers have similar surface energies, which helps stabilize assembled microdomains, but the lack of difference in surface energy negatively impacts the driving force for separation into sufficiently ordered patterns[40]. This self-contradiction is a key problem in polymer coblock development.

---

[37] R. Farrell *et al.*, "Large-scale parallel arrays of silicon nanowires *via* block copolymer directed self-assembly," *Nanoscale*, pp. 3228-3236, Apr., 2012.

[38] G. Liu, *et al.*, "Integration of Density Multiplication in the Formation of Device-Oriented Structures by Directed Assembly of Block Copolymer–Homopolymer Blends," *Adv. Funct. Mater.*, vol. 20, pp. 1251-1257, Apr., 2012.

[39] R. Gronheid, *et al.*, "Defect reduction and defect stability in IMEC's 14nm half-pitch chemo-epitaxy DSA flow," *Proc. SPIE*, vol. 9049, doi: 10.1117/12.2047265, 2012.

[40] C.J. Hawker and T.P. Russell, "Block Copolymer Lithography: Merging "Bottom- Up" with "Top-Down" Processes," *MRS Bulletin*, vol. 30, pp. 952-966, December 2005.

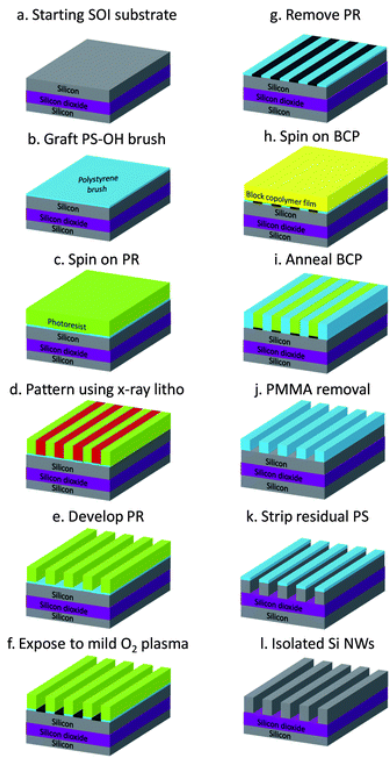

Figure 4: Chemoepitaxy approach to forming nanowires by DSA.

The spacing of diblock copolymers is proportional to the product of Flory-Huggins interaction parameter χ and the degree of polymerization of the polymer chain n, where χ is inversely proportional to the solubility of a binary mixture of the components[41]. Block copolymers force two immiscible monomers to covalently bond in a diblock chain. Self-assembly occurs in part because polymer block A repels polymer block B. Because the monomers are covalently attached, their ability to separate is limited by the length of the polymer chain. Thus the size of the chain, which is proportional to n, determines the minimum feature size. Smaller monomers lead to smaller polymers which lead to tighter pitches on patterned structures. But, reducing χn also reduces the driving force for separation. This in turn reduces the crispness of line edges and introduces more pattern faults. Process development will continue to focus on polymer synthesis that promotes covalent bonding of smaller, immiscible polymers. However, decreasing the solubility of the copolymer monomers tends to induce larger differences in the surface energy of a given block copolymer, leading to a skinning phenomenon where surface forces prevent

[41] A. Chremos, *et al.*, "Flory-Huggins parameter χ, from binary mixtures of Lennard-Jones particles to block copolymer melts", *J. Chem. Phys.*, vol. 140, doi:10.1063/1.4863331, 2014.

pattern transfer[42]. Modification of the DSA block copolymer stack with inert interstitial layers and tailored anneals may mitigate this problem[43].

Defectivity is a consideration as well. DSA is relatively immature as a process technology and understanding of the root causes and solutions for defects is still evolving[44]. Figure 5 shows a rogue's gallery of typical DSA patterning defects. A complicating factor for DSA is that it results in a block copolymer mask with very little contrast between the constituent polymers, which makes the structures difficult to image with a scanning electron microscope (feature sizes are far too small to be visible optically). Research into defect metrology is on-going and advances in defect metrology might speed implementation of DSA, but it is less likely to advance the technical capability. However, DSA has its own unique patterning defect mode with what are termed segregation defects, where the coblock polymer does not have energy to form ordered arrays. This defect mode leads to patterning errors, and the incidence of occurance tends to increase as feature size decreases. So, in that sense, advances in defect metrology that mitigate segregation defects do have the potential to advance DSA capability overall.

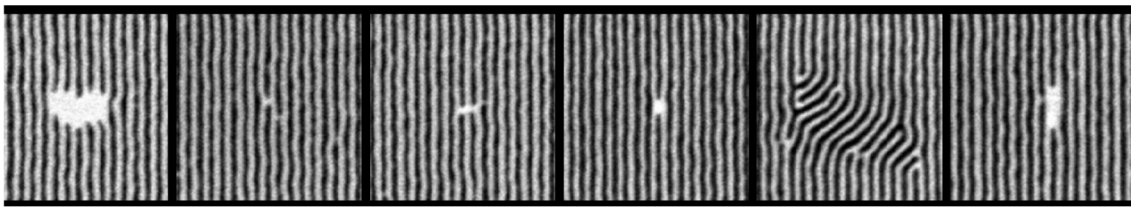

Figure 5: SEM images show defects termed darkspot, bridge, multibridge, particle dislocation cluster, and collapse.[45]

---

[42] R. Lawson, *et al.*, "Simulation study of the effect of differences in block energy and density on the self-assembly of block copolymers," *Proc. SPIE*, vol. 9049, doi: 10.1117/12.2046603 , 2014.

[43] J. Zhang, *et al.*,"New Materials for Directed Self-Assembly for Advanced Patterning," *Proc. SPIE*, vol. 9051, doi:10.1117/12.2046328, 2014.

[44] R. Gronheid, *et al.*, "Defect reduction and defect stability in IMEC's 14nm half-pitch chemo-epitaxy DSA flow," *Proc. SPIE*, vol. 9049, doi: 10.1117/12.2047265, 2012.

[45] Ibid.

5. Channel Engineering

Overview

Transistors have long used trace amounts of dopants to change the electrical characteristics of silicon. Typically the source and drain have different dopant profiles than the channel under the gate. The speed with which carriers can traverse the channel helps determine the rated speed of the integrated circuit. Faster is better. To make faster ICs, the standard approach has been to shrink the linear dimensions of the transistor to reduce the diffusion distance for carriers. However, to manage power density on increasingly transistor-dense ICs, the industry reduced the supply voltage which works against speed. Since the early 2000's, to improve speed while also reducing power, the industry has used strained channels by epitaxially depositing germanium on silicon. Single crystal Ge has a lattice constant that is slightly larger than that for Si, which causes strain in the Si crystal. The result is higher mobility of carriers in the channel, and thus higher speed. An example of a strained channel under the gate of an Intel IC is shown in the TEM in Figure 6[46]. As device dimensions continue to shrink, the speed advantages of inducing strain in a silicon channel have diminished and now the industry is considering replacing the silicon channel altogether[47].

Technology Challenges

Replacing the channel is expected to be complicated, even by microfabrication standards. Germanium and III-V semiconductors are the likely candidates, but highly manufacturable and cost-effective approaches to integrating those materials on Si and then subsequently forming gate dielectrics and silicides remains to be demonstrated. For the positive-channel field effect transistor (pFET), which carries holes across the channel, germanium transports charge four times faster than silicon. For the negative-channel FET (nFET), which depends on the movement of electrons, III-V materials are the focus. Indium gallium arsenide (InGaAs) in particular is of interest because it has electron mobility more than six times that of silicon.

---

[46] S. Thompson, *et al.*, "A 90 nm logic technology featuring 50 nm strained silicon channel transistors, 7 layers of Cu interconnects, low k ILD, and 1 um^2 SRAM Cell," *Int. El. Devices Meet. Tech. Dig*, pp. 61-64, Dec., 2002.

[47] R. Stevenson, "Changing the transistor channel," *IEEE Spectrum*, http://spectrum.ieee.org/semiconductors/design/changing-the-transistor-channel, 2013.

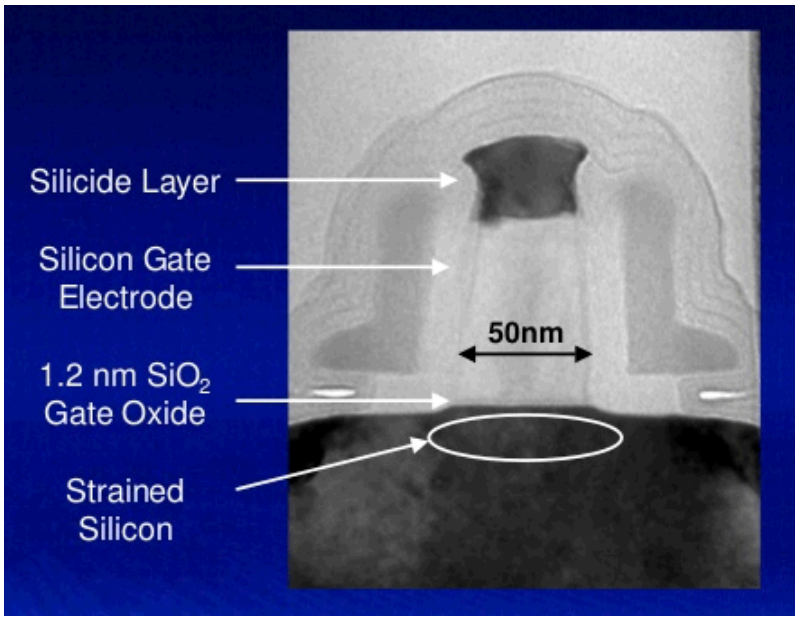

Figure 6: TEM of the strained channel on an Intel IC.

Among the possible approaches to instantiate alternative channels are selective epitaxy and blanket epitaxy[48], condensation[49], melt growth[50], and wafer bonding[51]. Epitaxy is slow, but it is a relatively clean vacuum process and relatively easy to control. Condensation, melt growth, and wafer bonding are all being investigated, but many challenges remain. Condensation is a variation on the epitaxial approach. Ge is grown on the single crystal silicon template and then diffused into the lattice with an anneal. This is repeated until the desired stoichiometry and film properties are achieved. Melt growth uses deposition of amorphous germanium followed by pulsed laser ablation to allow the germanium to anneal with crystalline order. A wafer bonding approach that results in Ge and InGaAs transistor channels has been demonstrated, but on germanium wafers. Our experience with wafer bonding informs our thinking that a wafer bonding approach will be difficult to make manufacturable due to challenges with topography, defects, and thermal cycles. We also note that state of the art integrated circuits are fabricated on 300

---

[48] IMEC press release on channel engineering, http://www2.imec.be/be_en/press/imec-news/imeciiivfinfet.html

[49] K. J. Chui, *et al.*, "Source/Drain Germanium Condensation for P-Channel Strained Ultra-Thin Body Transistors," *Int. El. Devices Meet. Tech. Dig*, pp. 493-496, Dec., 2005.

[50] M. Voelskow, *et al.* "Buried melting in germanium implanted silicon by millisecond flash lamp annealing," *Appl. Phys. Lett.*, vol. 93, doi: 10.1063/1.2993332, 2008.

[51] M. Yokoyama, *et al.*, "III–V/Ge High Mobility Channel Integration of InGaAs n-Channel and Ge p-Channel Metal–Oxide–Semiconductor Field-Effect Transistors with Self-Aligned Ni-Based Metal Source/Drain Using Direct Wafer Bonding" *Appl. Phys. Express*, vol. 5, doi:10.1143/APEX.5.076501, 2012.

mm Si wafers, but there are no device quality 300 mm Ge or InGaAs wafers to bond to silicon wafers. As a consequence, any wafer bonding approach will use only a fraction of the surface of a 300 mm Si wafer, and all subsequent processes will have to compensate for the topography inherent to the step edge from the submounted wafer.

Current State of the Art

Selective epitaxy appears to be the frontrunner. With selective epitaxy, a dielectric material covers silicon. A small, tapered window is opened, thus exposing a small area of silicon. Single crystal Ge or InGaAs are epitaxially grown in those areas, using the crystal lattice of silicon as the template. Because the window is small, defects attributable to lattice constant mismatch (line and stacking defects) are mitigated. Because the sidewalls are sloped, a crystal of usable size is grown. Any overburden in crystal size is presumably planarized back using CMP, though that detail is not covered in the literature to our knowledge (we have demonstrated a similar process in our fab). Once the crystal is formed, building the transistor is relatively straightforward. The epitaxial approach is being explored for both pFETs and nFETs, but the integration for nFETs is more complicated. For pFETs, pure germanium or germanium with silicon is a tractable integration. For nFETs, in November 2013 IMEC showed some details of their approach in a press release[52]. The IMEC process uses a layer of epitaxially grown germainum on silicon, followed by indium phosphide on Ge, followed by epitaxially grown InGaAs on the InP. A TEM of the resulting structure is shown in Figure 7. The driver for this complex integration is minimizing leakage currents. IMEC grows the thin film of Ge and then anneals it to create an ordered surface on which a higher quality InP film can be grown. InP crytal growth in a necked trench tends to terminate threading dislocations in the crystal lattice at the walls of the trench[53]. By this manner, the top of the InP crystal is free and defects and amenable to growing a high quality InGaAs crysta. Without this complex integration, lattice irregularities in the Si surface (which was exposed to plasma etch and wet processes) translate into the InGaAs, leading to In-In and Ga-Ga shorts that eliminate the semiconductor properties of the III-V crystal.

---

[52] IMEC press release on channel engineering, http://www2.imec.be/be_en/press/imec-news/imeciiivfinfet.html

[53] C. Merckling, *et al.*, "Selective area growth of InP in shallow trench isolation on large scale Si(001) wafer using defect confinement technique," *J. Appl. Phys.*, vol. 114, doi: 10.1063/1.4815959, 2013.

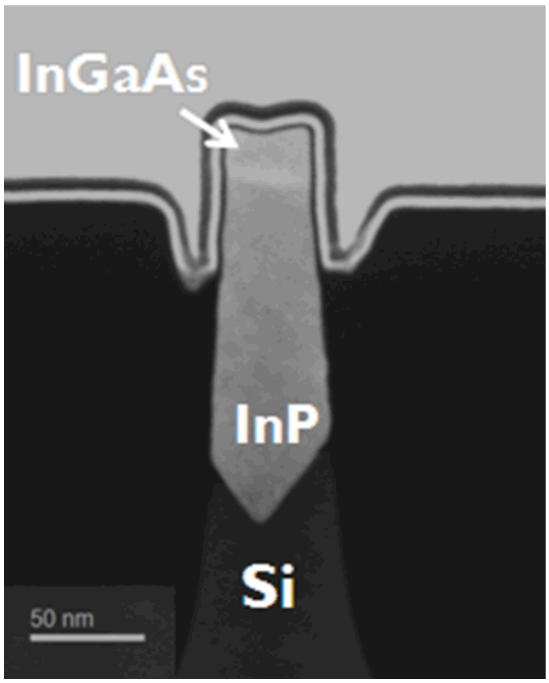

Figure 7: TEM of IMEC's instantiation of an InGaAs channel on Si substrate.

Challenges and Future Directions

Variations of selective epitaxy are likely to stay at the forefront of channel replacement. In fact, it has been reported that Sematech abandoned its blanket epitaxy approach to focus on selective epitaxy[54], thus concentrating more resources on the integration of selective epitaxy. However, problems remain. Epitaxial crystal growth is a slow process and throughput for high volume production remains a concern (throughput is less of a concern for niche markets that do not require high volumes of parts). Growth rates can be accelerated but the quality of the resulting crystal is compromised, leading to leaky transistors. Research that improves crystal growth rate while preserving the quality of the crystal will be important. Crystal quality drives the IMEC approach of growing multiple stacks to enable a final high quality InGaAs crystal, but that is a complex and expensive piece of integration. Research that improves the quality of the base lattice template without multistack integration will also be important.

[54] R. Stevenson, "Changing the transistor channel," *IEEE Spectrum*, http://spectrum.ieee.org/semiconductors/design/changing-the-transistor-channel, 2013.

6. Memory Integration

Overview

For decades, computer main memory architecture has been relatively static, seeing only incremental improvements in performance and density. Technological and economic factors are pushing memory systems to become more architecturally diverse. In this section we open with an overview of the DDR standard and new technologies that have enabled new innovations, and then we discuss the three major potential sources for technical surprise: Multi-Level Memory (MLM), Procesing-In-Memory (PIM), and advanced packaging techniques (Chiplets and 3D Integration).

Unlike a conventional DDR DRAM-based memory system, MLM combines several memory technologies together to achieve a "best of all worlds" memory system. MLM could lead to the rapid development of computers with much better cost/performance profiles. Though programming challenges abound, there is a strong potential to build a memory system that offers several times more effective bandwidth and costs less than half of a conventional (DDR) system.

PIM attempts to remove the "Memory Wall" by putting processing closer to the memory. There is some reason to believe that a PIM system can be implemented without dramatic change to the software. But, PIM offers a spectrum of implementation options, most of which are largely unexplored.

Chiplets & 3D stacking provide new techniques and design flows to integrate processors, memory, and specialized logic. This could lead to an overall decrease in the device's cost with little or no performance or power impact. More importantly, these techniques could lead to tech surprise through the speedy creation of more efficient, specialized, devices that can outperform conventional general purpose monolithic designs.

Architecture - A Post-DDR World

We define computer main memory as memory that is directly accessible from the processor with a load or store instruction, as opposed to storage that must be accessed through an IO call (e.g. the file system). Main memory for supercomputers, desktops, and most embedded systems is dominated by Dynamic Random Access Memory (DRAM). DRAM provides relatively low cost and power compared to SRAM memory and much lower latency and higher bandwidth than hard disks or tape. For the last 20 years, Double Data Rate (DDR) DRAM has been the most common form of DRAM memory. Technological and economic factors are moving which will break this dominance.

The DDR standards (DDR and its successors, DDR2, DDR3, and DDR4) are defined by JEDEC[55], an industry group comprising all of the major memory and processor venders. From 1989 to today, JEDEC standards (JESD21-C 2015) have defined the increasingly

[55] Formerly the Joint Electron Device Engineering Council, now just JEDEC

complex electrical interfaces that allow a processor to connect to DDR memory (JEDEC 2012). The DDR-based memory standards share a number of commonalities[56]:

- To save cost, they have a relatively small number of IO pins per DRAM chip
- To increase bandwidth, multiple chips are ganged together into a package (generally a DIMM), forming a wide parallel bus
- The individual memory chips are "dumb" – they contain minimal logic and are controlled in a master-slave fashion by the processor
- The interface allows little room for interpretation or differentiation – encouraging DDR memory to be a simple commodity[57].

These commonalties have benefited the computer industry by creating a common standard for memory that has seen steady performance increases, but it has hindered architectural innovation. Because it is a largely undifferentiated commodity, DRAM manufacturers have slim margins. This has led to consolidation in the industry: in the 1980s over a dozen companies produced DRAM, today, with the recent Micron/Elpida merger, roughly 90% of DRAM manufacturing is controlled by Samsung, Hynix, and Micron[58]. The low margins and intense competitiveness in the industry[59] have historically left little room for architectural innovation. Even with support from major manufacturer Intel, competitors to DDR, such as Rambus or FB-DIMMs, have failed due to the economies of scale and the low cost of commodity DDR.

Things are changing. There is an emerging consensus that there will be no DDR-5 standard.[60],[61] The increasing bandwidth and capacity requirements of future systems and the technical challenges of scaling the wide parallel DDR interconnect are too great. Main Memory architectures will become more diverse. Advances in packaging, growth in non-volatile memories, and possibilities for merging processing and memory will lead to major changes in how memory is designed and integrated into systems. Additionally, memory vendors see new architectures as a way to differentiate themselves and offer higher-margin products. Rather than "racing to the bottom" to provide a low cost "dumb" commodity, memory vendors are seeking to create advanced products which can claim a larger portion of computer revenue. These new architectural combinations open up possibilities and could lead to tech surprise.

---

[56] B. Jacob, *et al.*; "Overview of DRAMS." In *Memory Systems: Cache, DRAM, Disk*, by Bruce Jacob, Spencer W Ng, and David T Wang, 315-341. Boston, MA: Morgan Kaufmann, 2008.

[57] DRAM prices are tracked and traded like other undifferentiated commodities like oil or grain: http://www.dramexchange.com/

[58] B. Gain; "Times Are A-Changin' in the DRAM Market." *EBN*, January 9, 2014.

[59] J. Kang; *A Study of the DRAM Industry.* Master's Thesis, MIT Sloan School of Management, MIT, Boston: MIT, 2001.

[60] E. Sperling; *Will There Be A DDR5?* October 10, 2012. http://semiengineering.com/will-there-be-a-ddr5/

[61] J. Handy; "Where are DRAM Interfaces Headed?" *EE Times*, March 18, 2014.

Many of these new approaches are enabled by 3D Integration with Thru-Silicon Vias (TSVs), which is an advanced packaging technique that allows much greater memory density within a single package. It also has the potential to reduce power consumption and increase bandwidth by replacing lengthy and limited off-chip wires with much smaller connections through the Silicon. 3D TSVs are a key enabler for other technologies like MLM and PIM. There is no single JEDEC standard for 3D memory, but several competing standards (See Table 1) are emerging:

- **Hybrid Memory Cube (HMC)**: Led by Micron, the HMC Consortium[62] is manufacturing an aggressive 3D integrated stack of memory. The HMC offers extremely high bandwidths (10-15x conventional DDR) and lower power consumption by switching from a wide parallel interconnect to a more efficient narrow high-speed serial connection. Additionally, the HMC combines the memory chips with a logic layer that provides resiliency features, simple atomic operations in memory, and simple routing to allow multiple HMCs to be chained together for more capacity. The HMC will be used in some network equipment, an upcoming Fujitsu supercomputer, and Intel's Knights Landing processor.
- **High Bandwidth Memory (HBM)**: Proposed by Hynix, the HBM memory is a JEDEC standard[63]**.** Like the HMC, HBM is comprised of a stack of memory. However, it does not include a logic layer of its own and instead uses a series of wide parallel busses to connect to a processor. It is geared towards short distance interconnections, such as over a silicon interposer. This limits the capacity of an HBM memory system. HBM bandwidth is lower than HMCs, but still much higher than conventional DDR.
- **Wide IO**: Another JEDEC standard, Wide IO and WideIO2 are 3D stacked memories aimed at mobile graphics applications.[64] As such, Wide IO emphasizes size, cost, and power.

These new memories are not only technically interesting, but represent a major shift in the memory industry. Instead of undifferentiated standardized commodity memory, the DRAM market is becoming more diverse and opening up opportunities for innovation and differentiation.

---

[62] Altera Corporation, ARM Ltd., International Business Machines Corporation, Micron Technology, Inc., Open- Silicon, Inc., Samsung Electronics Co., Ltd., SK Hynix, Inc., and Xilinx, Inc., http://www.hybridmemorycube.org/

[63] JEDEC. *HIGH BANDWIDTH MEMORY (HBM) DRAM.* Standard, Arlington: JEDEC, 2013.

[64] JEDEC. *Wide I/O 2 (WideIO2).* Standard, Arlington: JEDEC, 2014.

Multi-Level Memory (MLM)

One possible future for memory systems is Multi-Level Memory (MLM). In MLM systems the main memory is comprised of two or more types of memory instead of a conventional DDR-DRAM-only main memory. By combining different memory technologies, an MLM system can potentially offer an order of magnitude more usable bandwidth and several times the capacity for a similar cost as a conventional memory system. However, substantial programming challenges must be overcome to realize this potential. Several vendors have announced (publicly or under NDA) future products that will use MLM or a roadmap which may contain MLM systems:

- Marvell has proposed a three level system combining some sort of High Speed DRAM (e.g. HBM, HMC, or WideIO) for performance with conventional DDR-DRAM and non-volatile Flash memory for capacity.[65]
- Intel's Knights Landing processor will use a modified HMC for performance plus conventional DDR DRAM for capacity.
- AMD is actively exploring 2- and 3-tiered memory systems.

Economic Impacts

Replacing DDR main memory with a combination of memory technologies may allow a memory system that provides high bandwidth and high capacity at low cost. Application analysis (See 'Algorithm and Software Impacts' below) indicates that, for many applications, a relatively small percentage of main memory accounts for most of the cache misses. If this portion was stored in fast, expensive (e.g. 3D stacked) memory and the bulk of data kept in slower, cheaper (e.g. DDR or Flash[66]) devices, it may be possible to realize the "best of both worlds."

---

[65] R. Merritt, "Marvell Shakes up SoCs, DRAMs." *EE Times*, February 23, 2015.

[66] K. Sudan, A. Badam, and D. Nellans. *NAND-Flash: Fast Storage or Slow Memory?* Tech Report, Department of Computer Science, University of Utah, Salt Lake City: University of Utah, 2012.

Table 1: Emerging Memory Comparision

Memory Technology	Interface, Packaging, Device	Relative Cost Per Bit	Typical Bandwidth / device
DDR4 DRAM	**Parallel multi-package DRAM**	**1**	**17 GB/sec (DIMM Module)**
Flash Memory	**Multi-package Non-Volatile**	**0.15**	**2-3 GB/sec**
HMC	**Serial 3D Stacked DRAM**	**3-5?[67]**	**160-240 GB/sec**
HBM	**Parallel 3D Stacked DRAM**	**2-4?[67]**	**128 GB/sec**
Wide IO	**Parallel 3D Stacked DRAM**	**3?**	**51 GB/sec**

Because several HMC and HBM memories are not yet in volume production, their price estimates are difficult. However, it is quite certain that Non-volatile Flash memory will continue to be much cheaper (per bit) than DDR-based DRAM and that HMC and HBM will provide substantial improvements in bandwidth per dollar.

Algorithm and Software Impacts

The largest impediment to successful deployment of MLM technology is software and algorithms. Software will need to be modified to place commonly used data in the fast "near" memory and less frequently used data in the "far" slow memory. Alternately, operating system or runtime algorithms will have to be implemented which transparently move data from one memory to the other by predicting future application requirements.

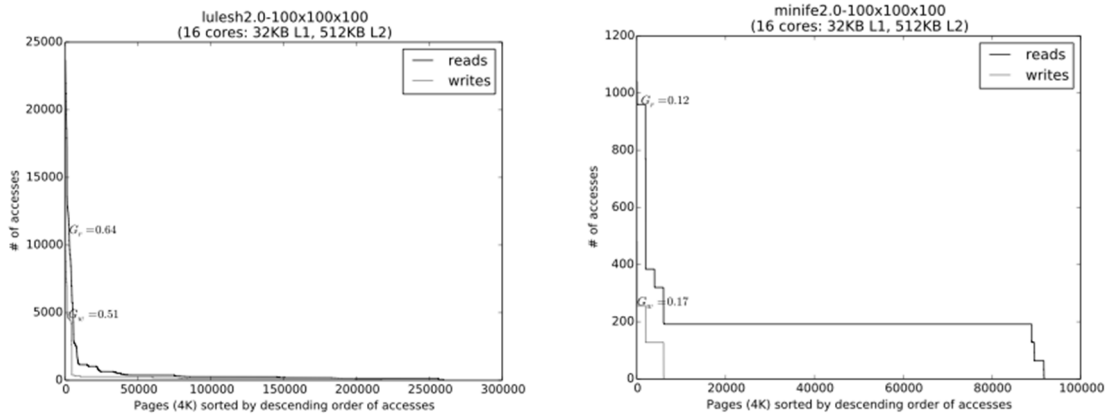

Figure 8 Application Main Memory Access Histograms

[67] As these products have not started shipping in quantity, prices are only estimates.

Some preliminary application analysis performed at Sandia indicates that application's interactions with main memory vary greatly. Figure 8 shows histograms of how often pages in main memory are accessed by the processor[68]. By sorting the histogram bins by frequency of access, distinctive patterns emerge. On the left, Lulesh, a hydrodynamics application, shows a very unequal distribution of memory accesses. Less than 5% of pages in main memory account for more than half of the memory accesses and an additional 15% of pages account for the vast bulk of memory requests. In contrast, Minife, a finite element application, has a more equal distribution. Though a small number of pages still account for a disproportionate number of requests, the bulk of the memory footprint still receives a number of accesses.

As a rough estimate of the level of inequality in memory accesses we have computed the Gini coefficient[69] for several applications. In a selection of DOE miniapps we have obsevered Gini coefficients as high as 0.92 (for reads in rsbench) and as low as 0.12 (for reads in minife)[70]. This leads to the understanding that there is no single one-size-fits-all solution for MLM data placement.

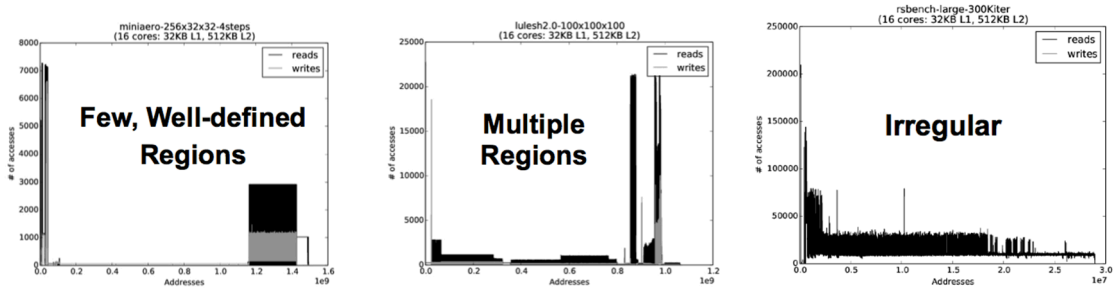

Figure 9: Memory Accesses by address

Figure 9 shows access histograms for different applications sorted by address. Like Figure 8, applications show considerable diversity. On the left, Miniaero, a 2D incompressible flow Navier-Stokes application, shows large, well-defined regions of memory that are accessed a similar number of times. This indicates that these codes could

---

[68] Note: the histograms only show post-cache accesses, better reflecting how often main memory is actually used.

[69] Originally used in economics to study income or wealth inequality, the Gini coefficient is a measure of statistical dispersion. A perfectly equal distribution (i.e. everyone has the same income or all memory pages area accessed the same number of times) results in a Gini coefficient of 0. A perfectly unequal distribution (i.e. only one memory page is ever accessed) would have a Gini coefficient of 1.0.

[70] Because the Gini coefficient was originally used as an estimate of income inequality in a society, this analysis allows us to proclaim that codes like Lulesh look like Botswana, while minife is more like Sweden and rsbench's write pattern like an oppressive medieval fiefdom.

be easily modified to identify which portions could fit in faster memory and which may be safely relegated to slower memory. In contrast, codes like Lulesh (middle), show multiple regions that may be more difficult to track down. Other codes, like Rsbench (right), a molecular dynamics code, have considerable variation in the number of accesses per page. These codes may be very difficult to perform *a priori* data placement on and may require more adaptive application or runtime methods to move data during execution.

Similar analysis by AMD has also found that the choice of which objects should be placed in slow or fast memory is often non-intuitive, even for expert programmers who are very familiar with a code. Because main memory only sees post-cache accesses, data structures that are accessed very frequently often do not account for many main memory requests because they are quickly loaded into cache. As a result, programmers will need to identify objects that are used "a lot, but not too much" and prioritize them against other similar objects. This is a difficult and complex task, which will require new analysis tools to help the programmer.

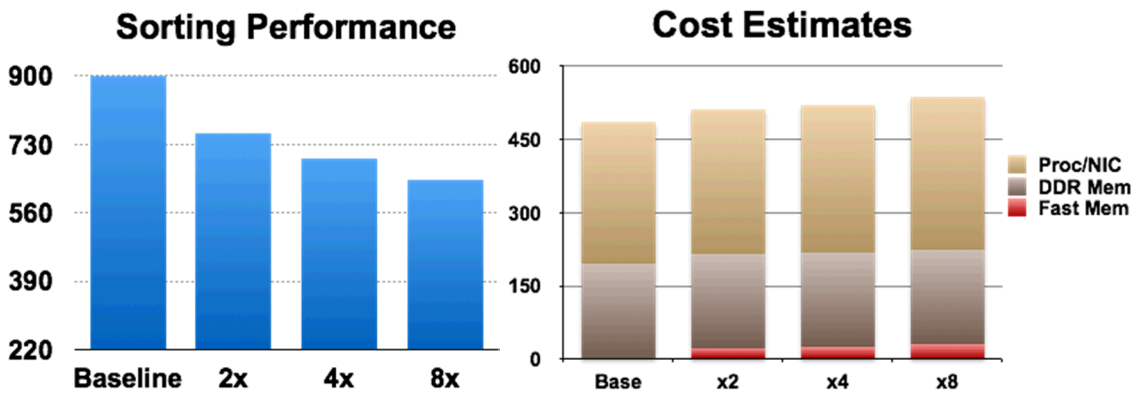

Figure 10: Performance and Cost Estimates for a MLM-Aware Sorting Algorithm

Though the software task is a difficult one, the potential for MLM is very high. Figure 10 shows the results of an experiment performed at Sandia where a simple integer sorting algorithm was rewritten to take advantage of two levels of memory, one similar to conventional DDR memory plus a faster stacked memory. The chart on the left shows the performance impact of different assumptions of the bandwidth increase for the additional high speed memory. The chart on the right shows cost estimates for a typical HPC node. This experiment demonstrated a conservative 40% performance improvement for only a 10% estimated cost increase. These results are conservative for a number of reasons:

- We assumed no overlap between memory transfer and the sorting. Future implementations could overlap this copying time with the computation.
- As seen from Table 1, 8x is a conservative multiplier of bandwidth for some of the emerging memory technologies
- The cost estimate assumed that the high-speed memory augmented the DDR instead of replacing a portion of it. If the high-speed memory replaced a portion

of the DDR, it could decrease the cost of the DDR and simplify the processor by switching from an expensive parallel wide interface to a cheaper serial interface.

Potential for Tech Surprise

HPC machines are often limited by the cost of the memory capacity the applications requre. HPC performance is often limited by memory bandwidth. With mono-level DDR main memories, the capacity and bandwidth are closely linked and limited. A fundamental change in this equation, provided by MLM, could lead to the rapid development of supercomputers with much better cost/performance profiles. An adversary who can more quickly adapt their software or who is simply not constrained by legacy software could rapidly deploy an effective MLM system.

As a simple thought experiment, consider a computer with a multi-level memory hierarchy based on the technologies from Table 1 and optimized for an application like Lulesh (Figure 8, left). The 5% of memory pages that dominate memory accesses could be placed in a small HMC-like memory sized to fit. The additional 15% that accounts for the bulk of the remaining accesses could be placed in conventional, low-cost DDR, and the "long tail" of infrequently touched memory pages could be placed in non-volatile Flash.

Table 2: MLM Thought Experiment

Memory	Size	Relative Cost/Per bit	Relative Total Cost
HMC	**5%**	**3.0**	**0.15**
DDR	**15%**	**1.0**	**0.15**
Flash	**80%**	**0.15**	**0.12**
Total			**0.42**

Table 3: MLM Average Effective Bandwidth

Memory	% of Memory Accesses	Bandwidth	Avg. Effective BW
HMC	**50%**	**240 GB/s**	**120 GB/s**
DDR	**40%**	**17**	**6.8**
Flash	**10%**	**3**	**0.3**
Total			**127.1**

Such a memory system could cost less than half of a conventional DDR-only memory system (Table 2). Since the majority of memory accesses would go to the HMC, the overall bandwidth would be increased substantially (

Table 3), with a comparable performance increase. The bulk of remaining accesses would go to DDR, so their performance would be no worse than a conventional DDR-only system. The small portion of memory that would go to Flash would be slower, but with

intelligent prefetching or application modifications, this may not have a large impact. The end result is a memory system that offers several times more effective bandwidth and costs less than half of a conventional (DDR) system.

Processing-In-Memory (PIM)

Another potentially disruptive future for memory systems is Processing-In-Memory (PIM). The basic idea of PIM is simple: Remove the "Memory Wall" by putting processing closer to the memory.

By moving computation closer to the memory, PIM architecture can take advantage of the higher bandwidth, lower latency and lower energy costs enabled by physical proximity. PIM is not a new idea, but unfortunately it has a long history of failure, however, technological changes have opened new possibilities. Early implementations of PIM were created by instantiating memory and logic elements on the same silicon die. This was done both by adding logic to a DRAM die[71] as well as by tightly coupling logic to a large amount of SRAM[72]. Both of these approaches have severe limitations. When using SRAM as the memory technology, the result was both fast memory and logic, but with a very limited memory capacity. The DRAM approach allowed for high memory capacity, but the optimizations used for a DRAM fabrication process[73] are not conducive to the creation of high performance logic, so the processing speed suffered.

In order to counterbalance the limitations of the physical implementation options available, many PIMs incorporated novel processor architectures, such as programmable logic arrays[74] and massive multi-threading (Brockman, *et al.* 2003). While these approaches created architectures with high compute potential, their adoption was limited, partially because applications would have had to adopt radically new programming models.

With the advent of 3D stacking with through silicon vias (TSVs), the physical limitations of the past have been removed and high performance, high capacity memory can be integrated with high performance logic. This close integration enables much higher bandwidth as large, expensive IO pins can be replaced with smaller, denser, TSVs. These vias allow a much greater number of wires to move data from one layer to another. Additionally, the close connection allows much shorter connections (10s of microns instead of 10s of centimeters) between the processor and memory.

---

[71] D. Patterson, *et al.*, "A Case for Intelligent RAM." *IEEE Micro*, pp 34-44, Mar./Apr., 1997.

[72] P.M. Kogge, "EXECUBE-A New Architecture For Scaleble MPPs." *Parallel Processing*, vol. 1, pp. 77-84, Aug., 1994.

[73] For example, DRAM usually uses only a few (2-4) layers of metal interconnect to save cost, while a processor may use over a 10 layers.

[74] M. Oskin, F. Chong, and T. Sherwood, "Active Pages: A Computation Model for Intelligent Memory." *Proc. 25th Ann. Int. Symp. on Comput. Architecture*, pp. 192-203, 1998.

"Conventional" PIM

One possible instantiation of PIM would be to stack DRAM memory on top of a more or less conventional CPU or GPU[75]. In this architecture, traditional processing cores are incorporated into the logic layer of an HMC or a logic layer below an HBM. The logic layer is already required to provide an interface with the DRAM layers above, so it does not require a dramatic change to the packaging. Additionally, the logic layer of an HMC is not fabricated in the same process as a DRAM because it contains complex logic and high-speed communication drivers. Thus, adding processing elements to the logic layer will not require dramatic changes to the fabrication process or compromise the capabilities of the processor.

Intel, AMD, and IBM are all exploring stacked memory with attached processing elements, either CPUs, GPUs, or dedicated vector processors. Sandia has also carried out experiments to test a "naïve" implementation of a PIM with conventional processors.

In the Sandia experiments a conventional multi-core CPU and memory system was augmented with stacked memories that contained processing elements consisting of similar cores. In simulation, the PIM cores ran at half the speed of the conventional CPU cores and had half the maximum issue bandwidth, making them roughly one quarter as powerful on a core-to-core basis. The PIM cores and the CPU cores shared the same address space and were kept cache coherent, with the PIM memory stacks containing the directory controllers. From a programming perspective, this looks like a mildly heterogeneous[76] multi-socket NUMA system – any core can access any piece of memory in the node, but more distant memory will take longer to access.

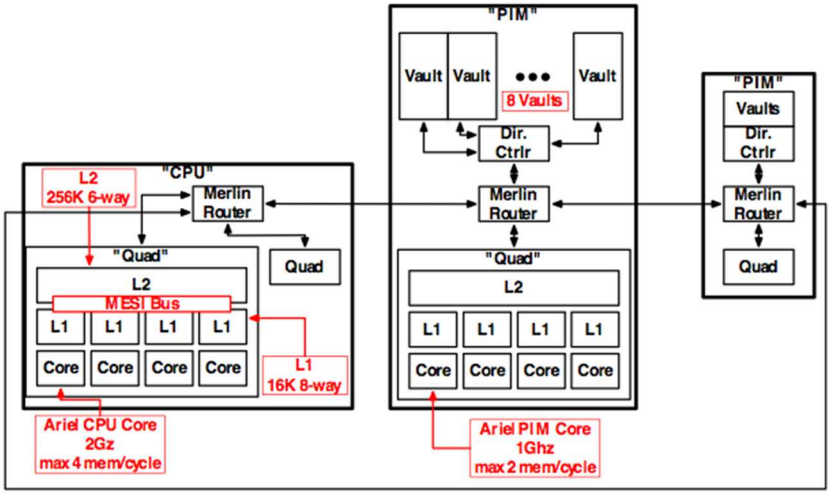

Figure 11: Sandia PIM Simulation Setup

---

[75] D.P. Zhang, *et al.*, "A new perspective on processing-in-memory architecture design." *MSPC '13 Proceedings of the ACM SIGPLAN Workshop on Memory Systems Performance and Correctness*, DOI: 10.1145/2492408.2492418, 2013.

[76] I.e. the processors have different speeds and issue widths, but use the same ISA.

This system has the benefit of minimal programmer impact – applications can certainly be optimized to take advantage of close memory, just like a conventional multi-socket system today, but it is not necessary to radically change the program just to get it to run or to achieve reasonable performance. Simulations (Figure 12) indicated that this naïve setup could achieve substantial performance gains. The figure below shows normalized execution time compared to the CPU-cores only case, with bars < 1.0 indicating a performance improvement. These results indicated that for a wide range of applications, including irregular memory accesses (GUPS), regular accesses (Stream), and scientific applications (MiniFE, Lulesh), substantial performance improvements could be realized. The one exception (Pathfinder, a small graph search code) showed little performance change, most likely due to a small problem size that did not parallelize well and fits in cache.

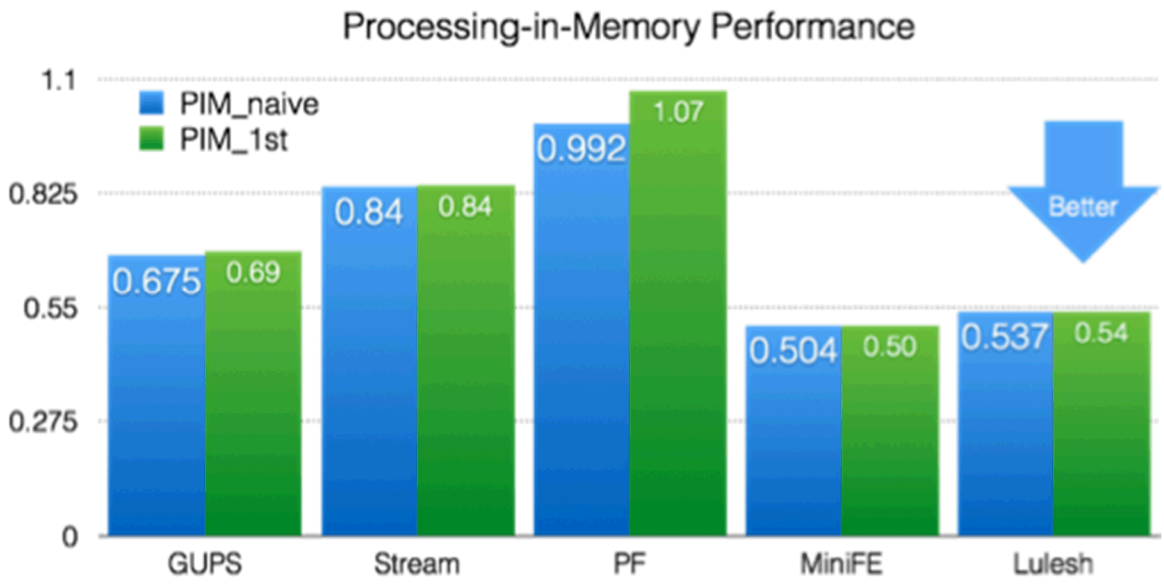

Figure 12: Simulated "Naive" PIM performance

We examined a very simple OS or runtime optimizations in an attempt to place pages closer to the cores that will use them (PIM_1st), but this did not have a dramatic impact on performance. Additionally, these experiments indicate that for most applications the increased parallelism of a PIM system does not dramatically increase the cache coherency traffic (Figure 13).

These findings give reason to believe that a PIM system can be implemented without dramatic change to the software. Additionally, advances in 3D stacking and memory make the realization of such a system much more possible.

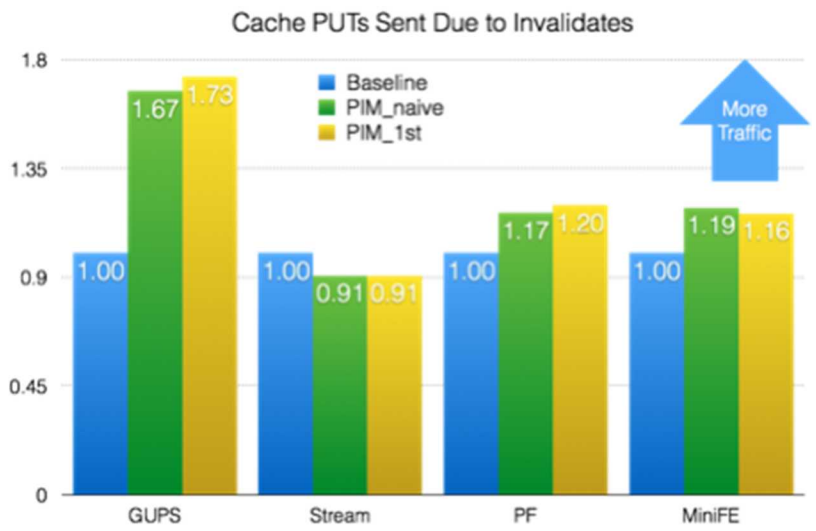

Figure 13: Naive PIM Cache Coherency Traffic

Architectures for Unstructured Data: Automata Processing

The regular grid-like structure of DRAM lends itself to massive parallelism. It may be possible to exploit this regularity to combine processing elements very closely to the memory system. Though no current mainstream products do this, there is some emerging work in this area.

One current non-Von Neumann architecture, and potentially disruptive new technology, is the Micron Automata Processor (AP)[77]. The AP is a custom designed memory built to support non-deterministic finite state automata at the hardware level. While geared towards complex regular expression matching[78] and non-deterministic finite automata it may be applicable to many other problem areas in cybersecurity and bioinformatics applications. Currently its potential is largely unexplored.

The automata processor is an adaptation of DRAM, which adds computing elements joined together with a routing matrix (Figure 14). Each column of DRAM (256-bits) contains a small piece of logic that can perform simple pattern matching. The results of this match are fed to a routing matrix that can combine results from different columns or determine which inputs to feed into the memory next. This input selector is based off of

---

[77] Micron Technology. *Automata Processing.* 2015. http://www.micron.com/about/innovations/automata-processing

[78] I. Roy and S. Aluru. "Finding motifs in biological sequences using the micron automata processor." *Proc. IEEE 28th Int. Parallel Distr. Process. Symp.,* DOI: 10.1109/IPDPS.2014.51, 2014.

the DRAM row selector that already exists in all DRAMs, leading to a very small and efficient compute specialized engine[79].

The current programming abstraction[80] for the AP favors defining finite automata, where states/nodes are joined by transition rules. Up to a maximum of 2304 transitions per node can be explored simultaneously, resulting in a massively parallel computation to discover strings accepted by the language.

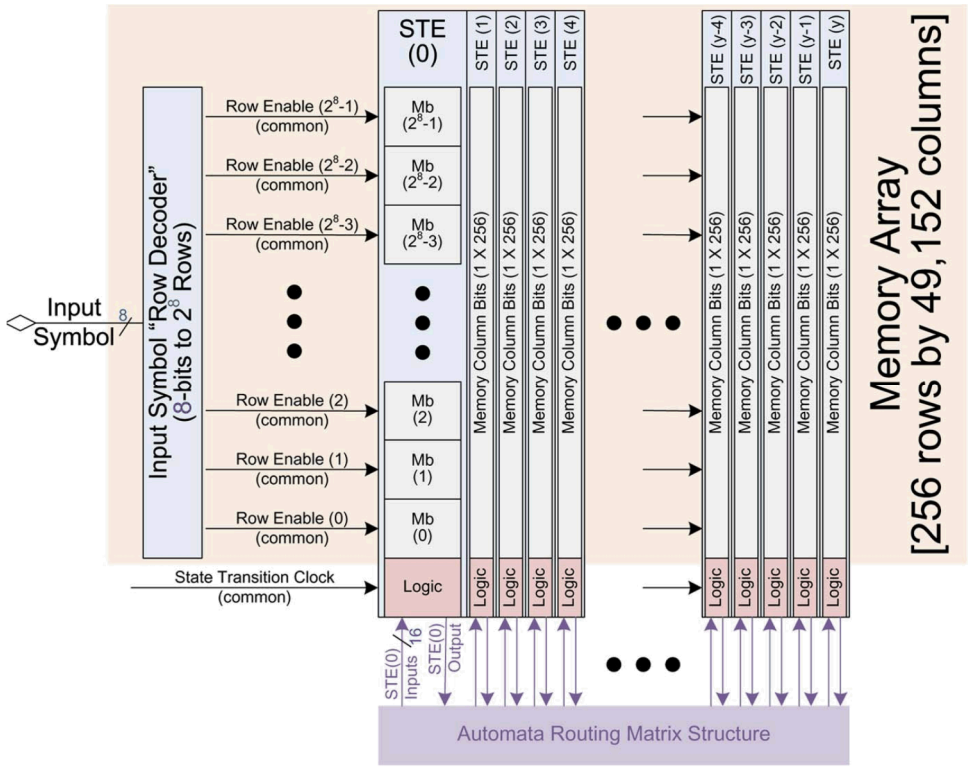

Figure 14: Micron Automata Processor Schematic

Potential for Tech Surprise

PIM offers a spectrum of implementation options, largely unexplored. Combine this with the diversity of the application space and there is a high possibility that a very high efficiency / high performance combination exists where a PIM solution can provide dramatically higher performance for a low cost or size/weight/power.

At present, there are several individual efforts, both in industry and research labs, to explore PIM implementations. However, for PIM to be successful commercially, it will require acceptance from memory, processor, system, and software makers. This will be a difficult task that is largely determined by non-technical factors such as business models

---

[79] P. Dlugosch, *et al.*, "An Efficient and Scalable Semiconductor Architecture for Parallel Automata Processing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 12, pp. 3088-3098, 2014.

[80] Accessed through a custom Micron-developed language

and IP rights. This bottleneck could become a window for a more nimble, integrated entity to produce a specialized proprietary PIM product and gain a substantial computational lead on the open market.

Design & Packaging

3D Stacking

From an architectural perspective 3D Stacking offers several key abilities:

- The ability to get a "best of both worlds" design for heterogeneous technologies. For example, it is key to enabling conventional PIM designs, stacked memory with high-speed serial interfaces, and future integration of silicon photonics.
- Reduced power and latency and increased bandwidth between chips. By moving from large, expensive chip-to-chip packaging towards small, efficient TSVs the physical distance between devices is reduced and one can run many more data paths between them.
- Business model changes. This level of integration opens the possibility for a part to have layers fabricated at multiple locations and then integrated. This could enable, for example, the integration of commercial parts with trusted parts with unique proprietary or classified capabilities.

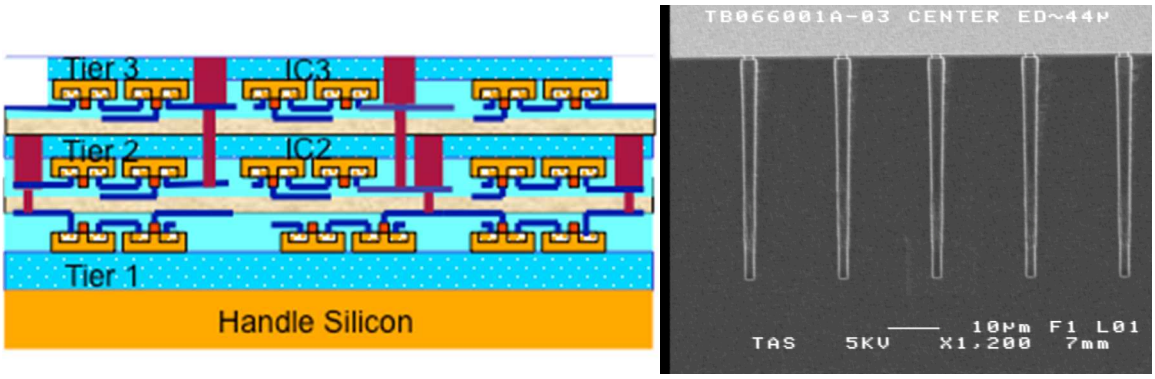

Figure 15: 3D stacking schematic (MIT) and TSV Micrograph (Sandia).

Chiplets

Another interesting development in packaging technology is the use of "chiplets" for prototyping or even production. The basic idea is to combine a series of small (<30 mm^2) chips together with 3D stacking or Silicon interposers to act as a large monolithic chip (>150 mm^2). Because the cost of a silicon chip increases superlinearly with its area, and because fabrication processes can be customized for different parts of a design, this could lead to an overall decrease in the device's cost with little or no performance or power impact.

The chiplet approach is not just a change in packaging, but a new approach to how electronics are designed. By moving from a 'monolithic' design process to a more modular approach, it is easier to "mix and match" existing components. In many ways this is a physical realization of the System on Chip (SoC) design flow that has come to dominate mobile devices. In both cases, existing components (possibly from different vendors) are used to rapidly build up a new design.

The cost savings of a chiplet approach may be between 2% and 16% depending on time frame and yield assumptions. However, the more important aspect may be the capabilities it enables:

- **Faster design**: Chiplets may allow new design elements to be integrated with existing IP blocks, allowing faster design turn around.
- **Flexibility**: Because the chiplet design flow can combine different fabrication processes (e.g. 28nm MPU with 14nm Memory with photonic) into a unified part, it could, like 3D stacking, enable a "best of all worlds" scenario.
- **Customization**: Because chiplets are physically integrated by packaging, it may be much easier to get a customized part than with a conventional monolithic production flow. For example, if a customer wanted a different ratio of CPUs to GPUs, a monolithic design would require a large design effort and the creation of new mask sets (costing several million dollars). A chiplet approach would still require some design work, but potentially much less.

- **Combining trusted & commodity**: An interesting possibility for national security applications would be the close integration of trusted components and untrusted or commodity components. Currently, monolithic production makes it difficult to integrate specialized functions with commercial parts in a trusted setting. With chiplets it would be possible to have a trusted design team and foundry produce a specialized function chiplet (e.g. cryptographic accelerator, special sensor, data movement engine, etc…) and integrate it closely with a high-performance commercial core (e.g. a GPU or CPU) without exposing the specialized functions to the open.

Potential for tech surprise

Advances in packaging offer considerable flexibility and have the potential to decrease design times. Technologies such as chiplets will enable the combination of existing IP blocks to be more efficient, and allow existing commercial IP blocks to be combined with proprietary fixed-function accelerators. This could lead to tech surprise through the speedy creation of more efficient, specialized, devices that can outperform conventional monolithic designs.

7. Optical Interconnects

Overview of Optical Interconnects in Computing

For the last 10 years or so, optical interconnects have been used to connect network nodes in high performance computing (HPC) systems. Typically, the technology is based on discrete transceivers that are electrically connected using printed circuit board traces to high-value electrical networking chips within the node (as in Figure 16). The data rates generally have followed the Ethernet data rate trend, with 10 Gb/s coming online after the year 2000 and 100 Gb/s coming online in 2010. The technology for high performance computing (HPC) interconnects lags the technology for data center interconnects, largely because the data center market is much larger. Generally, and especially for hyper-scale data centers like Google or Facebook, the needs are similar even if the applications are different. One trend that we are seeing is a bit of a 'slowdown' in cost-effectiveness of new technology; 100 Gb/s transceivers, even 5 years after their introduction in 2010, are too expensive, draw too much power, and are too large compared to what the vendors would ideally want.

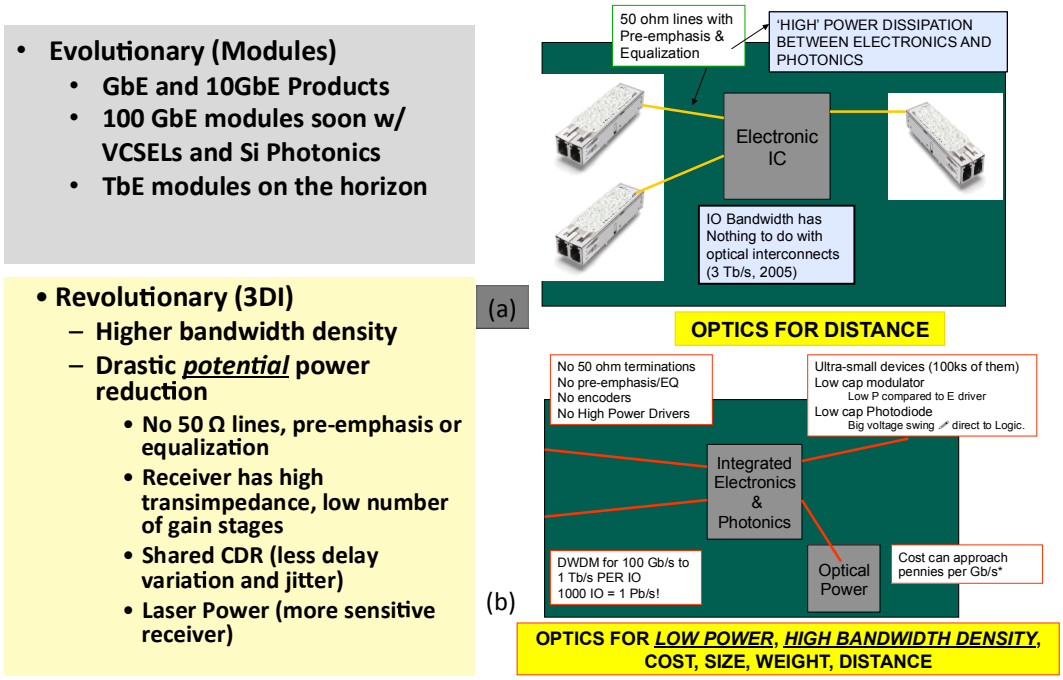

Figure 16: a) Traditional optical transceiver architecture b) Intimately integrated transceivers with high-value electronics ICs (processors, networking ICs, etc.)

In the longer term, a substantial power savings and interconnect bandwidth improvement can be realized using optical transceivers that are intimately integrated with the networking chips as shown in Figure 16(b). This approach avoids the bandwidth constriction of the electrical lines interconnecting the network chip(s) to the optical

transceivers and the power of the 50Ω line drivers and receivers of the short electrical interconnections shown in Figure 16(a). The integrated approach might be heterogeneously integrated, similar to the familiar electrical 3D chip-stacks, for example in hybrid memory cubes (HMC). Indeed, it is likely more feasible to consider an intimate hybrid integration as the photonics would not need to be compatible with ever-increasingly complex integrated circuit technology nodes. However, a hybrid approach would still allow electrical connections to the optical transceivers in the 10µm scale, yielding electrical connection energies in the range of 1fJ/bit for a 10µm line at 1V. By contrast, a 50Ω line driver at 10-25Gb/s is upwards of 1pJ/bit (1000X), even for short connections lass than 10cm in length. The other advantage of the integrated approach is that optical signals can be multiplexed on different wavelengths; for example 40 wavelengths at 25Gb/s is indeed 1Tb/s on a single optical path. Our internal calculations show that a complete optical energy consumption of 100fJ/bit is feasible, but there is a lot of engineering work to get there.

In this section, we will discuss the evolution of both approaches and highlight any potential breakthroughs that might lead to a tech-surprise compared to path pursued by the research community and market forces. In the 5-10 year time frame, the solutions will have a data rate of 1–4Tb/s, as roadmaps in the 2-3 year time frame are in place for early adopters of 400Gb/s transceivers. This analysis is by no means complete; the fields of optical interconnects and optical communications are so developed that it is difficult to include a treatment of all the potential technology surprises that might come to play over the next 5-10 years. Hopefully, this will give the reader an introduction so that they may request further discussion on items that seem to be of interest.

Current Optical Transceiver Technology

Multi Mode Solutions

Transceivers that use vertical cavity surface emitting laser (VCSEL) transmitters over multimode fiber are leading the market today. This is largely because the packaging and installation costs of using multimode fibers is lower than that using single mode fibers. A VCSEL has the advantage of producing a round beam that is easy to couple into the fiber compared to an edge-emitting laser. Currently, the small-form factor pluggable (SFP+) package with 10Gb/s data rate is quite prevalent. VCSELs with serial data rates up to 28Gb/s (25Gb/s plus error correction coding) are becoming prevalent, although not quite at the product stage. VCSELs operating to 56Gb/s using pre-emphasis and equalization have been demonstrated in a laboratory environment. Today, 100Gb/s may be realized using 4 independent links at 28Gb/s using either multi-core fiber or a fiber ribbon[81]. Transceivers designed to be interconnected with 4 to 12 fiber ribbons have been in production for several years, but multi-core fibers with many smaller cores that fit in the

---

[81] C. Cole, "Next generation 100G client optics," in *2011 37th European Conference and Exhibition on Optical Commun. (ECOC)*,Geneva, Switzerland, 2011.

cross-sectional area of a standard 125µm diameter fiber might be a lower-cost solution in the long run[82]. However, this technology still needs low cost connector development.

There are a couple methods that one can see to get to 100Gb/s per fiber core or per fiber (single core) using multimode technology. Research demonstrations of serial data transmission using pre-emphasis and equalization at 56Gb/s [83] and PAM-4 (4-level signaling) at 30 symbols per second or 60Gb/s[84][85] are the highest rates per fiber (core) reported. These might be extended to PAM-4 at 56Gs/s or PAM-8 at 35+Gs/s, where Gs is gigasymbol with 2bits/s and 3bits/s per symbol for PAM-4 and PAM-8 respectively. Second is to use dual-polarization[86], either in conjunction with PAM-4 at 28Gb/s or serial at 56Gb/s. Third is to use a 100Gb/s serial duo-binary transceiver, band-limited to about 40GHz. There have been limited demonstrations of coarse wavelength division multiplexing[87], but this approach does not seem to be gaining momentum because of the difficulty of providing wavelength multiplexing and demultiplexing solutions in a multimode environment.

Beyond 100Gb/s (and maybe beyond 56Gb/s), the solution is almost for sure going to require multi-core fiber or multiple fibers. A 16-fiber element active optical cable has been demonstrated at 25Gb/s per connection (400Gb/s aggregate)[88]. A Terabit solution would require either more cores or a fiber ribbon of multi-core fibers. The main technical challenge for these solutions to be adopted in the marketplace is the development of low-cost multi-core fibers, multi-core fiber connectors, and transceivers compatible with the multicore fibers. At the higher rates, these packaging and fiber challenges are increased with the need to provide up to 20 cores or a multiple-fiber ribbon where each fiber has multiple cores.

Single Mode Solutions

While HPC is not particularly cost sensitive, it does not have the development dollars to produce its own products and must rely heavily on those developed by the data center

---

[82] B.G. Lee, *et al.*, "End-to-End Multicore Multimode Fiber Optic Link Operating up to 120 Gb/s," *J. Lightwave Technology,* vol. 30, no. 6, pp.886-892, Mar. 2012.

[83] D. Kuchta, "A 56.1 Gb/s NRZ modulated 850 nm VCSEL-based optical link," *Proc. Optical Fiber Comm. Conf.*, Anaheim CA, 2013.

[84] N. Quadir, P. Ossieur, and P.D. Townsend, "A 56 Gb/s PAM-4 VCSEL driver circuit," *Proc. Irish Systems and Signals Conf.*, Maynooth Ireland, Jun 2012.

[85] K. Szczerba, P. Westbergh, and M. Karlsson, "60 Gbits error-free 4-PAM operation with 850 nm VCSEL," *Electron. Lett.*, vol. 49, no. 15, pp. 953-955, Jul 2013.

[86] D. V. Kiesewetter, "Polarisation characteristics of light from multimode optical fibres," *Quantum Electron,*, vol. 40, no. 6, pp. 519-524, 2010.

[87] R. Michalzik, 'Four-Channel Coarse WDM 40 Gb/s Transmission of Short-Wavelength VCSEL Signals Over High-Bandwidth Silica Multi-Mode Fiber," University of Ulm, Ulm Germany, Annual Report, pp. 53-58, 2000.

[88] http://investor.finisar.com/releasedetail.cfm?releaseid=831944

market. In the data center market, Google and Microsoft publically acknowledge that they are interested in single-mode solutions that are required for transmission at higher data rates (e. g. 25Gb/s) over distances greater than 100m[89]. While only a small percentage of links occur over those distances, they presumably want a solution that does not need fiber-replacement as data rates increase (like multiple cores/fibers would) and would like a uniform solution to handle short and long interconnects.

The primary benefit of single mode solutions is the ability to multiplex multiple wavelengths onto a single fiber using integrated optics solutions. Other benefits include the ability to manipulate the optical phase (required for coherent solutions) and the ability to lower the receiver power dissipation by virtue of the much smaller photodiode.

Silicon photonics has emerged as a leading platform to realize single mode interconnect products over the long term for data center applications. The advantages of silicon photonics compared to indium phosphide (InP) integrated optics are primarily cost, size, integration complexity and potentially yield; of course, lasers are still needed and those likely will be made in an InP platform and integrated with silicon. Several research groups have demonstrated laser integration with silicon in a way compatible with silicon photonics. The high index-contrast of silicon photonic waveguides allows one to realize more compact optical filters and complex optical circuits to be realized than would be the case in a low index-contrast platform.

Silicon Photonics can be broken down into two general classes of devices, resonant and non-resonant devices. The majority, if not the entirety, of companies developing silicon photonics products are using non-resonant devices. Within that category, there are currently three leading approaches to >100Gb/s transceivers. First have been demonstrations of 4 wavelengths at 25-28 Gb/s. These use Mach-Zehnder modulators and either ladder filters (based on passive interferometers) or Eschelle gratings. This approach is aimed at the data center market. The second use coherent communications techniques and modulate data in a dual-polarization, quadrature phase shift keyed (DP-QPSK) modulation format. Currently, this is aimed more at long-distance or metro-regional communications systems, but can be potentially used in data centers with some caveats (we will discuss below). Third is an approach using a new 64-element fiber array that uses multiple multimode fiber-optic cables, but with silicon photonics transceivers[90]. This is a product today with an aggregate bidirectional bandwidth of 800Gb/s.

Resonant Silicon Photonics to 400Gb/s and Beyond

By using dense wavelength division multiplexing (DWDM), 400Gb/s can be simply realized with 16 wavelengths at 25/28 Gb/s and 1Tb/s can be realized with 40 wavelengths. Conceptually, this is a simple solution. Many university research groups

---

[89] "OIDA Roadmap Report: Future Needs of Scale-Out Data Centers," Anaheim CA, Mar 2013.

[90] http://www.intel.com/content/www/us/en/research/intel-labs-silicon-photonics-mxc-connector.html

and a few largely government-funded commercial companies are looking into the use of micro-ring and micro-disk modulators and filters to implement DWDM in a low-cost, low-energy manner. The technology has several serious challenges related to the variations of dimension, material, and environmental factors, especially the refractive index. This manifests itself in variations in the DWDM optical multiplexer (filter) wavelengths and pass band shape and loss as well as modulator extinction ratio and loss that is a function of the variation in resonant wavelength as a function of these parameters. Closed loop control circuits have been demonstrated at several research laboratories to stabilize the resonant wavelength of one device[91], but for a DWDM system, one needs multiple circuits working concurrently. One also has to apply the technique to second order filters on the receiver side of a DWDM link to achieve adequate rejection of the adjacent wavelength channel, something that hasn't been demonstrated so far. Variations in coupling coefficients are a secondary effect, but need to be managed so as to not cause too much degradation in optical filter response and adjacent channel rejection.

Lastly, a cost-effective source of 16–40 wavelengths needs to be implemented. Heterogeneous integration of III-V materials with silicon has been demonstrated (and depending on the definition, is implemented in products). It remains to be seen if 40 lasers can be implemented at a low enough cost compared to a single laser with 40 fibers, but we expect substantial work in that direction. In addition, a silicon nitride ring can generate a 'comb' of laser lines that can be used as the 40 sources with a single laser 'pumping' as single integrated device. These devices are still in research phase and are not as efficient as individual lasers, but they might eventually offer the most cost effective solution overall.

Importantly, these are solvable problems, but they require investment. The commercial marketplace hasn't yet required 400Gb/s and 1Tb/s of bandwidth per fiber, so it is willing to accept non-DWDM solutions. But as aggregate data rates continue to grow almost unabated, the need for increased bandwidth between switches and routers in data centers will require solutions such as this.

Research Paths for Non-Resonant Optical Modulators Compatible With Silicon Photonics

A fairly simple analysis can show that approximately 0.25°C temperature control is needed for a 1dB power penalty out of a typical ring or disk modulator[92]. For a ring-based optical mux or demux, the tolerance can be much greater (>1 °C), and it depends on the channel spacing, acceptable crosstalk (power penalty), and acceptable pass band channel characteristics. For 400Gb/s and 1Tb/s applications, other more fabrication tolerant, but larger DWDM optical multiplexers can also be used (e.g. arrayed waveguide

---

[91] J. A. Cox, *et al.*, "Control of integrated micro-resonator wavelength via balanced homodyne locking," *Optics Express,* vol. 22, no. 9, pp. 11279–11289, 2014, and references therein.

[92] A. L. Lentine, *et al.*, "Electronic interfaces to silicon photonics," *Proc. SPIE,* **vol. 8989,** doi:10.1117/12.2045696, Mar 2014.

gratings). So, the modulator presents the greatest challenge in controlling its resonant wavelength, and if the research community can invent and demonstrate a broadband modulator with wide fabrication, temperature, and material property tolerances, then the adaptation of DWDM silicon photonics will be greatly accelerated. This section discusses approaches to a suitable optical modulator.

Germanium-Based Electro-Absorption Modulators

Electro-absorption modulators on silicon have been demonstrated in germanium and silicon germanium[93][94][95]. These have been made with small sizes and energy consumption below 1fJ/bit, although that particular modulator still needs further optimization of its optical characteristics. High-speed modulation, while not demonstrated per-se on this particular modulator, is easily attainable, based on the physics of the device. A modulator such as this can modulate tens of channels over a fairly large temperature range (40°C) with a 1dB power penalty[96].

The largest challenge is the low-cost, high-yield integration of the silicon germanium modulators with the silicon photonics platform. Many groups, including ours at Sandia National Laboratories, have an integrated germanium detector in their silicon photonics platform. Lately, there have been sporadic publications in the area of germanium modulators on silicon, so it could be considered a technological surprise if someone developed the platform for use for product development.

Indium-Tin-Oxide (ITO) Modulators

There have been a few demonstrations of modulators based on indium-tin-oxide[97] [98]. The operation of the modulator requires a huge carrier concentration that alters the absorption of the modulator. The modulator is very simple to make in theory, but in practice requires a great-deal of material development. All modulators made to date have been extremely slow; this might be due to either the design of the device or the quality of the material. It

---

[93]Y. H. Kuo, *et al.*, "Quantum-confined stark effect in Ge/SiGe quantum wells on Si for optical modulators," *IEEE J. Sel. Topics Quantum Electron.*, vol. 12, no. 6, pp. 1503-1513, 2006.

[94] S. Ren, *et al.*, "Ge/SiGe quantum well waveguide modulator monolithically integrated with SOI waveguides*," IEEE Photon. Technol. Lett.*, vol. 24, pp. 461–463, Mar. 2012.

[95] J. F. Liu, *et al.*, "Design of monolithically integrated GeSi electro-absorption modulators and photodetectors on an SOI platform," *Optics Express*, vol . 15, no. 2, 2007.

[96] R. D. Kekatpure and A L. Lentine, "The suitability of SiGe multiple quantum well modulators for short reach DWDM optical interconnects," *Optics Express* , vol. 21, no. 5, pp. 5318-5331, 2013.

[97] R.T. Chen, *et al.*, "Indium tin oxide single-mode waveguide modulator," *Proc. SPIE,* vol. 1583, 1991.

[98] V. J. Sorger, D. Kimura, R.-M. Ma, and X. Zhang, "Ultra-compact silicon nanophotonic modulator with broadband response," *Nanophotonics*, vol. 1, no. 1, pp. 17–22, Jul. 2012.

would be surprising to see this device become a part of a silicon photonics platform, because of the material difficulties.

Graphene-Based Modulators

There have been demonstrations of graphene-based modulators using similar structures to the ITO devices[99]. Graphene is still in its infancy, so it also has a number of material challenges. Although many of the published papers claim high-speed devices, they are for the most part slow, with the fastest published modulators in the 1-2Gb/s range. We expect that the prognosis for improving graphene-based modulators is better than that of the ITO devices, but again, there isn't much published research in this area at present.

III-V Devices on Silicon

III-V devices are very similar in operation to germanium electro-optical modulators. They were first demonstrated about 20 years ago, and were developed in a surface normal platform to a fairly high degree of integration with 4000 -32000 IO at 850 nm using GaAs/AlGaAs multiple quantum well (MQW) modulators[100][101]. InP-based modulators at 1.55µm have been developed using heterogeneous integration with a silicon waveguide platform, but the evanescent coupling of those devices precludes small, low energy devices[102]. However, the platform could be modified to produce efficient silicon photonics with efficient III-V modulators operating at 1.55µm. It is a more difficult integration challenge than using germanium for about the same level of performance, and even more challenging, is that it is likely to be a much higher cost option. Thus, we think it is highly unlikely that industry will follow this approach.

III-V Photonic Integrated Circuits (PICs)

Following the discussion of III-V devices on silicon, III-V integrated optics circuits not integrated with silicon have been commercialized as transmitters and receivers on metro-regional and long distance communications by integrating complete DWDM transmitters on a single photonic integrated circuit[103]. These devices not only include multichannel modulators, but also multiple lasers, semiconductor amplifiers, and channel combiners

---

[99] M. Lui, *et al.*, "A graphene based broadband optical modulator," *Nature*, vol. 474, pp. 64-67, May 2011.

[100] A.L. Lentine, *et al.*, "High-speed optoelectronic VLSI switching chip with >4000 optical I/O based on flip-chip bonding of MQW modulators and detectors to silicon CMOS," *IEEE J. Sel. Top. Quantum Electron.,* vol. 2, no. 1, pp.77-84, Apr 1996.

[101] T.L. Worchesky, *et al.*, "Large Arrays of Spatial Light Modulators Hybridized to Silicon Integrated Circuits," *Appl. Optics*, vol. 35, no. 8, pp. 1180-1186, 1996.

[102] Y.-H. Kuo, H.-W. Chen, and J. E. Bowers, "High speed hybrid silicon evanescent electro-absorption modulator," *Opt. Exp.,* vol. 16, no. 13, pp. 9936–9941, Jun. 2008.

[103] S. Hurtt, *et al.*, "The first commercial large-scale InP photonic integrated circuits: current status and performance," *2007 65th Ann. Dev. Res. Conference,* Notre Dame IN, pp.183, Jun 2007.

(which are passive on the transmitter, rather than optical multiplexoers, but gain makes up the extra loss). However, owing to the relatively large size of the InP modulators and the low refractive index contrast of the InP material system, the devices are rather large. Hence, while all the performance and function for integrated 400Gb/s and >1Tb/s transmitters and receivers is available in these platforms, their power dissipation, physical size, and cost make them a poor choice for data center and computer optical interconnects. Thus, we do not expect to see them ubiquitously applied in such applications.

Slow-Light Mach-Zehnder Modulators

Mach-Zehnder modulators are the mainstay of today's optical high-speed optical communications devices. The main advantage of the devices is that the optical bandwidth can be fairly high (30-40nm), meaning it is relatively insensitive to device fabrication, material properties, and temperature, although a closed loop control circuit is still generally used. Devices operating in excess of 50Gb/s have been demonstrated[104], with commercial products in the 28Gb/s (100GbE) to 40Gb/s readily available. Early devices and even some of today's systems use lithium niobate discrete devices, which have good power handling and ideal sinusoidal response. InP devices are used in several products; they strike a balance between small form factor and good performance. Silicon Mach-Zehnder modualtors today are the device of choice for low-cost applications, and device performance (speed, voltage) is approaching that of the InP devices. However, all the Mach-Zehnder devices are large (0.5-5mm), consume relatively high energy (1–10pJ/bit), and have fairly high optical losses (typical ~ 4dB in the high-state), especially for long devices with low voltage drives. (Note: the drive voltage is inversely proportional to length). However, if the light can be slowed through the use of photonic crystals or some other method, the device can be made much shorter (tens of nm) with a corresponding reduction in energy consumption[105]. The challenges of slow-light modulators include making uniform slow-light material (this is generally a section of waveguide with a very-small regular array of holes), and achieving sufficiently high optical bandwidth is also a challenge, as slow light is usually narrow band. If the latter challenge isn't solved, then one might as well use resonant modulators that are reasonably well-developed.

Electro-Optic Modulators (organic polymers, lithium niobate)

The main point of using a different material is to realize a stronger change in refractive index with voltage compared to silicon. This might also achieve less temperature dependence, depending on the materials and design. The materials might also achieve a change in absorption with voltage, further enhancing modulation depth. Other potential advantages include higher power handling and lower loss as silicon suffers from two-

---

[104] http://www.intel.com/content/dam/www/public/us/en/documents/intel-research/Intel_SiliconPhotonics50gLink_FINAL.pdf

[105] S. Akiyama, *et al.*, "A 1V peak-to-peak driven 10-Gbps slow-light silicon Mach-Zehnder modulator using cascaded ring resonators," *Appl. Phys. Express* , vol. 3, no. 7, doi:10.1143/APEX.3.072202, 2010.

photon absorption that limits both loss and power handling. Lithium niobate has very low loss, but its electro-optic coefficient (change in index with voltage) is not going to allow significantly smaller and lower energy devices. However, organic polymers might enable smaller and lower energy devices and faster devices[106]; to date they suffer from repeatability/reproducibility problems and have un-proven reliability. However, a small amount of research continues to be done in this area, so maybe some day this will help improve modulation on silicon.

Another approach is to design a compound silicon/polymer waveguide structure that has a composite refractive index sensitivity to temperature that is greatly reduced from silicon alone[107]. These structures would allow us to use resonant optical modulators and optical filters with characteristics that don't change with temperature. However, there still will be variations from manufacturing. Post-tuning of the refractive index is an option to compensate for variations in manufacturing[108]. Then, the only question is whether the resonant shift variations that occur from two-photon absorption will be large enough to cause significant modulator increases in loss or decreases in extinction ratio. An approach like this requires a stabilized laser source, because there is now no method to track the incoming laser source. A silicon photonics ring or disk modulator or filter with a feedback control loop, could track any variations in the laser wavelength, so it would not need to be stabilized over time and temperature. That notwithstanding, this is a promising alternative to silicon micro-rings and micro-disk modulators and filters, but trades complexity in feedback control circuits in the silicon case for complexity in manufacturing in the polymer resonant device case. The potential benefit is greater modulation as a function of voltage.

DWDM single mode VCSELs and nano-lasers

If we can replace a silicon photonics DWDM modulator with an active single mode DWDM laser, then the power efficiency might improve and the complexity of the chips might be reduced, as the laser source for a modulator system is eliminated. However, there are several drawbacks to the approach. A single mode VCSEL 1550 nm transceiver is available commercially, but that is only a single non-integrated device. The VCSEL and silicon photonics would still need to be integrated. The downside of the approach is that the integration of a VCSEL with silicon photonics or some other optical platform that provides optical mux and demux. Second, the lasers must be either temperature controlled or the optical mux and demux must track the lasers. Because the lasers in this application are not an optical power source but a data source, the lasers must be

---

[106] L. Alloatti, *et al.*, "100 GHz silicon–organic hybrid modulator," *Light: Science & Applications*, vol. 3, e173, doi:10.1038/lsa.2014.54, 2014.

[107] V. Raghunathan, *et al.*, "Athermal Silicon Ring Resonators," *Integrated Photonics Research, Silicon and Nanophotonics and Photonics in Switching*, Monterey CA, doi:10.1364/IPRSN.2010.IMC5, Jul 2010.

[108] D.A. Bender, *et al.*, "Precision laser annealing of silicon devices for enhanced electro-optic performance," *Proc. SPIE*, vol. 8967, doi: 10.1117/12.2037339, Mar 2014.

modulated. Unlike a modulator that can be driven arbitrarily to a high and low state, a laser has a slow turn-on and needs to be biased precisely at threshold for high-speed operation. The feedback control circuit to do this across temperature, where the laser threshold and slope efficiencies both change, is more complex than that needed to control the resonant frequency of a micro-disk modulator.

Nanolasers are smaller form factor single mode lasers that might be more easily integrated with a waveguide platform, such as a silicon photonics platform. However, the devices are still in their infancy. The principal challenge is in reducing the cavity losses so that room temperature lasing can occur. At the current state of the art, the devices are so inefficient that they heat up too much. Hence, most experiments are performed under pulsed conditions or with optical pumping, where electrical dopants and contacts that increase optical losses are not present.

Methods of Increasing the Data Rate Using Multiplexing

Wavelength Division Multiplexing

Dense and coarse wavelength division multiplexing has been used in telecommunications systems for more than 20 years. However, it hasn't really found its way into data center communications and local area networks yet. The primary reason has been cost. The cost to integrate 4 lasers in a package with an optical mux and demux has been too expensive compared just running 4 times as many fibers. But as density demands increase and silicon photonics, heterogeneous integration, and related technologies reduce the cost and difficulty of implementing wavelength division multiplexing it will ultimately beat multiple fiber solutions in cost. Initially, coarse wavelength division multiplexing, with wavelength channels that are spaced by 20nm with uncooled sources that are allowed to drift will be the first implementations. But as integrated optics, especially silicon photonics and the heterogeneous integration of silicon photonics with III-V lasers, continues to mature, we will see more DWDM entering the marketplace. The real crystal ball question is the degree to which other multiplexing techniques will be used in conjunction with DWDM, and the baud rate and bit rate of the individual channels. In other words, will it be simpler to merely double the bit rate or double the number of channels? History suggests the former, but it has a more than linear economic cost above some bit rate that today is about 25Gb/s. This section addresses how to increase that bit-rate while keeping the overall frequency response of the electronics (i. e. baud-rate) below some nominal value (today 25Gb/s).

Multi Level Signaling

We briefly discussed using multilevel signaling earlier. Research demonstrations of both optical and electrical signaling have been done at 50-60Gb/s using PAM-4 (pulse amplitude modulation with 4 levels)[109] [110] and very recently using PAM-8 at 105Gb/s [111].

---

[109] N. Quadir, P. Ossieur, and P.D. Townsend, "A 56 Gb/s PAM-4 VCSEL driver circuit," *Proc. Irish Systems and Signals Conf.*, Maynooth Ireland, Jun 2012.

In these cases the symbol rate is 25–30Gs/s and 35Gs/s respectively, where Gs/s is giga*symbols* per second. Clearly, 100Gb/s data transmission per wavelength is feasible using PAM-4 or PAM-8 in the not too distant future; the main question is will it be cost-effective with other methods of achieving increased bit-rate per wavelength.

Coherent Optical Modulation Formats

Long distance 100GbE systems all use a higher modulation format; in this case quadrature-phase shift keying (QPSK). Each symbol is transmitted with one of four potential phase states; hence there are two bits of information encoded in each symbol. Two of these signals are transmitted concurrently, one on each polarization. So, each 'symbol' is transmitted at 25Gb/s (28Gb/s with forward error correction), with two polarizations and two bits per 4 phases amounting for 4 bits transmitted for each symbol.

200Gb/s modulation has been demonstrated by using the same equipment to transmit both amplitude and phase levels (quadrature-amplitude-modulation or QAM) versus just phase levels. 16-QAM, with 4 levels on each of two orthogonal phases contains 4 bits per symbol; with dual polarization, now 8 bits per symbol or 200Gb/s bit rate at 25Gs/s symbol rate is transmitted.

These systems require coherent detection between an optical local oscillator and the incoming optical signal on the receiver side. In these metro-regional and long haul communications systems, orthogonal very high-speed analog to digital converters (ADCs) convert the signals in alternate phase axes to a sampled stream. A high speed, high-performance digital signal processor (DSP) samples these data streams and provides a 'real-time' clock recovery by pattern matching to the overhead bit patterns in the optical transport network (OTN) data frame. The DSP also performs enhanced forward error correction (FEC) and dispersion compensation.

Using this sort of DSP in a HPC or data center optical interconnect application is a non-starter for both power and cost reasons. However, an optical phase lock loop (PLL) could be used in place of the DSP for the coherent phase recovery at the minimal cost of an extra laser (not insignificant, but at most a 2X factor) and some electronics. There is little dispersion in a short reach such as a data center, and even less in an HPC application, and thus no reason to compensate for dispersion. In an interconnect application, we certainly would not need advanced FEC and may be able to eliminate FEC entirely. For a QPSK implementation, we would not need the ADCs, because there are only two levels per phase quadrature, although we would need them for a QAM system at 200Gb/s. But still, a short-reach system with a higher modulation format would have drastically reduced

---

[110] K. Szczerba, P. Westbergh, and M. Karlsson, "60 Gbits error-free 4-PAM operation with 850 nm VCSEL," *Electron. Lett.*, vol. 49, no. 15, pp. 953-955, Jul 2013.

[111] B. Gomez Saavedra, *et al.*, "First 105 Gb/s Low Power, Small Footprint PAM-8 Impedance-engineered TOSA with InP MZ-Modulator and Customized Driver IC using Predistortion," *Optical Fiber Comm. Conference*, Los Angeles CA, doi:10.1364/OFC.2015.Th4E.4, Mar 2015.

power and footprint relative to a telecommunications application. We have not seen anyone proposing this architecture, but they ought to be.

With the ability to get to 100–200Gb/s per wavelength, we could envision 1Tb/s interconnects in a coarse wavelength division multiplexed arrangement, where the wavelength tolerances are much relaxed relative to a dense wavelength system. We still believe the DWDM system has the potential for lower power, but a CWDM/coherent system might have a shorter road to implementation, yet still pack all the data on a single fiber pair.

Multicore Fibers

Multicore fiber merely multiplexes signals in space versus wavelength or polarization. System demonstrations have already been performed with both single mode and multimode multicore fibers. Thi solution is in the long run more expensive – because the fiber costs more and that will not come down in cost in a manner similar to microelectronics and micro-photonics. However, we are very likely to see these as intermediate solutions until DWDM is cost effective and suffieciently developed for short reach applications.

Mode-Multiplexing

Mode multiplexing and angular momentum multiplexing (a form of mode multiplexing) offer another orthogonal axis to increase data rate. As multiple signals are launched into a coupled many-core fiber or multimode fiber, the signals mix and are at first glance indistinguishable at the fiber output. However, with digital signal processing, we can separate the mode-mixed signals into their original form.

If the system uses a coupled few core fiber solution, then the DSP should merely be able to solve a system of linear equations to arrive at the individual bit streams. However, in a true multimode fiber, more complex processing is needed to construct the original signals. Nonetheless, the development of an optical spatial equalizer to undo the mode mixing could in theory eliminate the DSP or at least reduce its complexity. Unless this can be implemented in standard multicore fiber, we don't see big advantage in mode multiplexing over transmission by a (mode-separated) multi-core fiber.

<u>Ultra-High Speed On-Off Keying (Digital) Modulation (>100Gb/s)</u>

As of this date, 25Gb/s (28Gb/s with FEC) offers the sweet spot for performance versus cost and complexity. Indeed, as previously noted, most 100Gb/s solutions are based on 4 bits per symbol, either using dual-polarization QPSK or 4 wavelengths at 25Gb/s for each wavelength. Several demonstrations for individual devices and basic links have been done at 50/56 Gb/s and a recent demonstration of a 100GHz bandwidth modulator has been done using organic polymers on silicon[112]. An earlier demonstration was done at

---

[112] S. Akiyama, *et al.*, "A 1V peak-to-peak driven 10-Gbps slow-light silicon Mach-Zehnder modulator using cascaded ring resonators," *Appl. Phys. Express* , vol. 3, no. 7, doi:10.1143/APEX.3.072202, 2010.

112Gb/s using band-limited components at 40Gb/s. In this demonstration, the band-limited nature of the components creates a duo-binary (3-level) signal[113]. A simple decoder, albeit at 112Gb/s, can decode the data.

The problems with the ultra-high speed solutions are many. First, the electronic interface circuits are terribly power hungry when the operation rate approaches the intrinsic device speed, and minimal parasitic capacitance will affect the circuit greatly. Second, the discontinuities from electrical packaging are increased proportionally. Third, the dispersion penalty for a given distance goes as the square of the bit-rate. For example, dispersion of standard single mode fiber (e.g. SMF-28) is 17ps/nm/km. (The reason it goes as the square of the bit-rate is that the dispersion allowed is reduced proportional to the bit rate and the optical bandwidth of the signal [in nm] is proportional to the bit-rate.) At 100Gb/s, the maximum fiber length will be limited to about 400m without some form of dispersion compensation. That is very close to the longest data center fiber distances. Thus, we believe that it is a reasonable assumption that maybe 50-56Gb/s might be the prevailing bit rate in 10 years, but probably not 100Gb/s.

Laser Source Integration with Silicon Photonics

Integrating low cost laser sources with silicon photonics in a low cost manner is a challenge, but at least a partially solved one. There are now several transceiver products with III-V integrated sources. However, the integration of 16–40 wavelength sources has not been successfully demonstrated. There are multiple approaches to this challenge.

First, heterogeneous integration (e.g. wafer bonding) of III-V materials to silicon can be used, with the lasing wavelength is determined in the III-V material or in the silicon material based on a frequency selective element such as a grating or micro-resonator. There is no technical reason that this cannot work but it is not a compact solution, as we need 16 – 40 individual lasers.

Second is a hybrid solution (i.e. flip-chip bonded). This is in fact a similar approach to wafer bonding, but perhaps more near-term as it can be executed using pick-and-place tools. While a detailed cost model for this approach has not been published, modern pick and place machines have huge throughputs and low costs (pennies per placement). The approach is helped if the lasers are nominally the same and the frequency dependent element is implemented in the silicon photonics chips. This is likely the most power efficient technique. We expect to see more research and developmentthat pursues this approach.

Third is a ‘comb’ source. This is the most compact and elegant solution. A single laser pumps a large resonator, often made of silicon nitride. The nitride converts the input laser line into a series of laser lines with resonant wavelengths spaced by the free spectral

---

[113] J. Lee, *et al.*, “Serial 103.125-gb/s transmission over 1 km ssmf for low-cost, short-reach optical interconnects,” *Optical Fiber Comm. Conference,* San Francisco CA, http://dx.doi.org/10.1364/OFC.2014.Th5A.5, Mar 2014.

range of the resonator. It is a challenge to achieve only the number of lasing lines needed without extras. There have been experiments showing that the noise characteristics of the laser lines do not degrade communications system margin compared to standard distributed feedback (DFB) lasers. Perhaps the greatest challenge in this technique is the integration of the silicon nitride that is suitable for a comb laser with a silicon photonics platform, although heterogeneous integration might be an alternate approach. This approach is the least power-efficient approach, because the nitride comb laser (conversion of the input line to a series of lines) has a decent but limited efficiency (20-40%) above and beyond that of the laser itself.

In any event, integrating a DWDM source with a silicon photonics platform is not an insurmountable challenge by any means. The largest question will be whether it can be done at a cost below that of multiple fibers and the degree to which the industry uses other forms of multiplexing first before turning to WDM.

System Packaging

We touched on the concept of 3D integrated photonics devices in Figure 16(b). It requires heterogeneously integrating photonics in high-value ICs rather than localized transceivers electrically connected to high-value ICs. Indeed this integration offers a substantial performance improvement and energy reduction. However, in addition to solving the heterogeneous integration challenges in a cost-effective manner, packaging is a huge challenge.

Today, optical fiber is used to connect racks of equipment with many high-speed interconnections. Fiber is the ideal medium as it is lightweight and flexible. Within a rack of equipment, there has been a lot of work on connecting printed circuit cards over a backplane, both in multimode and single mode configurations. As of this writing, these optical connectivity subsystems are available commercially, but are not often cost competitive with their electronic counterparts. However, they do offer a potential greater bandwidth density improvement, especially a single mode system that allows DWDM through its waveguide connections. Going to the next level down, there has been less work on directly interfacing the optics between ICs on a printed circuit card, even transceivers. The vision would be that the printed circuit cards would have both optical and electrical traces on them. Over distances within a small printed circuit card, it is much cheaper to electrically connect high-value chips to transceivers either on the front-plane or backplane. Electronic chips are almost always soldered to PCBs rather than socketed (though occasionally a processor is socketed) because the devices and connections are so reliable that when one fails, the entire PCB is replaced. Optical devices are almost always socketed as thus easily replaced, because their reliability isn't perceived to be as good. It is questionable whether standard pick and place machines and commercial PCB soldering techniques could adequately align a multi-mode optical guide on the PCB with the optical source or detector on a chip; there would be almost no chance to do this for a single mode waveguide, where the value in bandwidth density is much greater. Perhaps one needs to use active beam-steering techniques using micro-electromechanical systems (MEMs).

There is also an issue of design expertise and tool availability. There are more ASIC designers than there are optical system designers today. If we imagine that a heterogeneously integrated platform and the packaging infrastructure and accompanying reliability becomes available commercially, it will need all the tools and infrastructure for ASIC designers to be able to design the optical I/O the way they do electrical I/O today. Similarly, the PCB designers will need to be as familiar with optical routing as they are with electrical routing. This is not impossible, but it is likely a long-term endeavor that will require careful coordination

On-Chip Optical Interconnects

The limitations of electrical on-chip interconnections not scaling with integrated circuit technology nodes are well documented. The key issue is that as conductive lines become thinner, the resistance increases but the capacitance does not decrease very rapidly, because it is dominated by fringing capacitance (edge effect) rather than parallel plate capacitance. So, the capacitance of a metal line is stuck at about 0.1fF/um. This lack of scaling of capacitance means that as the technology node scales, the energy does not scale at a given length and the data rate transmitted through a thin line actually decreases at that given length. The electrical solution to the reduced bit-rate is to add buffers to the line, so that the interconnect length is shortened. That is, if the original design called for a 1mm line, we can use 10 line lengths of 100µm with buffers at the intermediate points. Re-timing along the way might be required depending on the detail of the design. However, while this solves the clock rate limitation, the energy is increased which leads to a situation where power density limits the bit rate, rather than the RC time constant.

So, while in theory intimately integrated optics as shown in Figure 16(b) can overcome both of these limitations (at a given crossover distance), a much lower cost needs to be achieved to start to replace electrical lines on chip with optics. Many of the technologies described, in particular silicon photonics resonant modulators and optical multiplexers, can be used for on-chip optical interconnects. For distances as short as 100µm, it can be more efficient to use optics[114]. However, it seems unlikely that the cost-point for photonic interconnections can be low enough in the near term (5 – 10 years) to enable the use of on-chip optical interconnections in mass-produced products such as computing and networking chips.

---

[114] E. Timerdogan, *et al.*, "A one femtojoule 'athermal' modulator," *Nature Comm.*, vol. 5, doi:10.1038/ncomms5008, Jun 2014.

Electrically Controlled Optical Switching Networks Vs. Optically Interconnected Electronic Switching Networks

There is a fair amount of research on optical networks within data centers and computers[115]. Optical switches of course have the advantage of being able to pass many channels of data (e. g. >1Tb/s per path) at low power consumption. However, data center and computing networks are different. In the data center application, there are long flows of data (elephant flows) that don't require fast switch reconfiguration and small amounts of data (mice flows) that require that the switches be reconfigured often. A key challenge is identifying the flows. Once that is done, large, low loss switches with microsecond reconfiguration times are ideal for data centers to route the elephant flows. However, in computing systems, short bursts of communication of a few bytes is much more common, so a slowly reconfiguring switch is less useful, although we don't believe the issue has received much research attention.

Another way to think about this is that a computing network has three components: (1) data-path routing, (2) data buffering, and (3) scheduling, path determination and arbitration. Electrically controlled optical routing only solves the first of the three areas – the one that is by far the simplest. So, it isn't obvious that integrated optics provides a large enough enhancement over optically interconnected electrical networks.

Optical Switching and Logic

Optical switching and logic can provide a new solution to all 3 of the components to computer networking. There is a very long history in optical logic with several books written in the field. A good but quick review of the advantages and disadvantages can be found in the paper by Miller[116]. In short, while optical effects are often very fast, qualitative aspects like having adequate signal gain, adequate I/O isolation, and low standby power make the prospect of a significant advance over electrical logic (perhaps with optical interconnects) extremely challenging.

One of the more interesting recent approaches is to use quantum phenomena[117]. This work is really in its infancy. The approach promises lower energy than electrical logic, which is in the vicinity of 1fJ/bit, but offers the potential for lower energy and faster

---

[115] D. Nikolova, *et al.*, "Scaling silicon photonic switch fabrics for data center interconnection networks," *Optics Express*, vol. 23, no. 2, pp. 1159–1175, 2015, and references therein.

[116] D.A.B. Miller, "Are optical transistors the next logical step?" *Nature Photonics*, vol. 4, no. 1, doi:10.1038/nphoton.2009.240, Jan 2010.

[117] Y.-D. Kwon, M.A. Armen, and H. Mabuchi, "Femtojoule-Scale All-Optical Latching and Modulation via Cavity Nonlinear Optics," *Phys. Rev. Lett.*, vol. 111, arXiv:1305.1077, Nov. 2013, and references contained therein.

speed. However, it will not be so easy to scale it to billions of gates to supplant CMOS, so this will require long term effort to find general applicability.

Conclusions

Optical interconnects are seeing more and more usage in high performance computing in addition to data center environments. We've highlighted where optical transceiver technology is going and might go in the near future. We've described some of the newer directions taking place in the research community and pointed out what we think are some potential technological surprises that might occur. Clearly, there is a lot of discussion in the community about silicon photonics which has advantages over other technologies by building on the decades long development of silicon CMOS. As this technology matures, it allows wavelength division multiplexing to be used, multiplying the bandwidth density of local interconnects the same way it did for long distance interconnects 20 years ago. While all the tools for utilizing silicon photonics for short-reach communications are here today, the next few years will continue to see inventions and innovations in silicon-compatible device technologies and sub-systems. The heterogeneous integration of silicon high value electronics with silicon CMOS will be a major influence on reducing the energy of high-performance computing and data center communications, but there are quite a few technological hurdles to be worked out. On-chip optical communications, widespread use of electrically controlled integrated optical (likely silicon photonics) switching networks, and optical switching and logic are likely beyond the 10 year time frame. The entire field has certainly moved forward at a rapid pace over the last 20 years, so it will be exciting to see the almost certain increase in use of optical interconnect for local communications in computing and data applications.

8. Neuro-Inspired Computing: A New Paradigm

Overview

Conventional computing has been benefiting from ever increasing performance levels provided by device scaling in CMOS technology for the last 40 years; however, this trend has reached the point of saturation due to physical and economic limits. All of our sensing, communication systems and data analysis approaches rely critically on the continual advances in scaling transistor size yet there are potentially only two more nodes left in conventional CMOS scaling, 10nm and 7/5nm, which are targeted for development and production by 2020.

There have been many attempts at finding a replacement device over the metal-oxide-semiconductor field effect transistor (MOSFET), like tunnel FETs[118] and carbon as the semiconductor (carbon nanotubes[119] and graphene[120], for instance). Even if these attempts provide attractive candidates for further development, the fundamental architecture underlying the computing capability remains the same: von Neumann/Turing architecture.

The highest payoff development direction may be a paradigm shift. Standard computing is based on algorithms and symbolic manipulation, but what if we can emulate the dynamic way neural systems represent, process, store and recall information in order to predict and control the physical environment? It is interesting to note that von Neumann and Turing both drew inspiration from the brain and made analogies to computing elements in their development of the electronic computer. Efforts that have taken place in following decades to replicate neural architectures in neural networks and attempts at developing artificial intelligence have met with limited success.[121] Despite the progress in deep learning networks (machine learning mostly based on algorithms around restricted Boltzmann machines[122]) which is enabled by the state of the art CMOS for computing (GPU clusters, accelerators, etc.), the underlying machinery still uses von Neumann/Turing architectures on MOSFET transistors.

---

[118] A. Seabaugh, "The tunneling transistor," *IEEE Spectrum*, vol. 50, pp. 35–62, Sep, 2013.

[119] M.M. Shulaker, *et al.*, "Carbon Nanotube Computer," *Nature*, vol. 501, pp. 526–530, Sept 2013.

[120] F. Schwierz, "Graphene Transistors," *Nature Nanotechnology*, vol. 5, pp. 487–496, May 2010.

[121] J. Schmidhuber, "Deep Learning in Neural Networks: An Overview," *Neural Networks*, vol. 61, pp. 85-117, Jan 2015.

[122] G.E. Hinton, *et al.*, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, pp. 1527-1554, 2006.

With the availability of high efficiency prediction and control systems, many critical applications could face a critical transition – what takes many kiloWatts of power and requires large footprint (many racks of electronics equipment) could be accomplished with a few tens of Watts, much smaller volume, and with a small fraction of the time of conventional computing approaches. Much more importantly, *neuro-inspired* computing systems could achieve much higher levels of functionality, such as situational awareness, capability to adapt on-line, one-shot learning and ability to generalize learned patterns from one system to another, completely unrelated system. These higher level functions will take large scale projects and dedicated resources to come to fruition. There are systems that are coming on line that will allow some of the early development directions to be evaluated.

The Neuro-Inspired Computational Elements (NICE) Workshop is held to provide a nucleation point for the development of the next generation of brain-inspired information processing and computation architectures that can process large, noisy, incomplete, "natural" data sets that do not lend themselves to convenient solutions by current systems. In the following section we summarize the 2015 NICE workshop, including a review of current applications, approaches, and resources; analysis which we characterize as workshop findings; and recommendations that are intended to steer the direction of research neuro-inspired computing. In the context of anticipating tech surprise, we view the NICE Workshop as a unique opportunity to leverage the combined expertise of an international assembly of experts to predict trends and developments in neuro-inspired computing.

2015 Neuro-Inspired Computational Elements (NICE) Workshop

The Neuro-Inspired Computational Elements (NICE) Workshop has been held for the last 3 years (2013, 2014, 2015) in Albuquerque, NM, with the goal of bringing together scientists, engineers and stake holders from a very wide spectrum—ranging from experimental neuroscience, computational neuroscience, algorithms, high performance computing, hardware and all the way to applications. In all technological development timelines, including microelectronics and computers, a critical need has driven and funded the early stages of development. In the case of microelectronics, it was the navigation and guidance needs of defense systems, while in the case of computers, it was the neutron diffusion simulations and weather prediction and forecasting needs that provided the "killer app." Such clear application cases are not currently defined for neuro-inspired and neuromorphic computing approaches, but this might become more obvious after some of the current projects gain more traction. The community that has convened at NICE workshops has seen and felt the need for a clear articulation of the "value proposition" of this new approach. As a result, there are significant efforts to provide the metrics, comparison cases and application examples to support further development of the science, technology and the ecosystem around this activity. We have succeeded in achieving the initial goal set out for the workshop, namely providing a environment for these discussions – and we plan on continuing with the workshop series to accelerate ongoing activity and support project and program development for neuro-

inspired and neuromorphic computing and providing the solutions for critical applications that might not be feasible within the current computing paradigm.

Workshop Technical Areas

During the presentations and discussions at the workshop, several application areas, distinct technical approaches and technology development paths were identified. Brief summaries of these are presented within Figure 17 below. Further detailed information (presentation recordings and presentation files) can be found at the event web site: **http://nice.sandia.gov**.

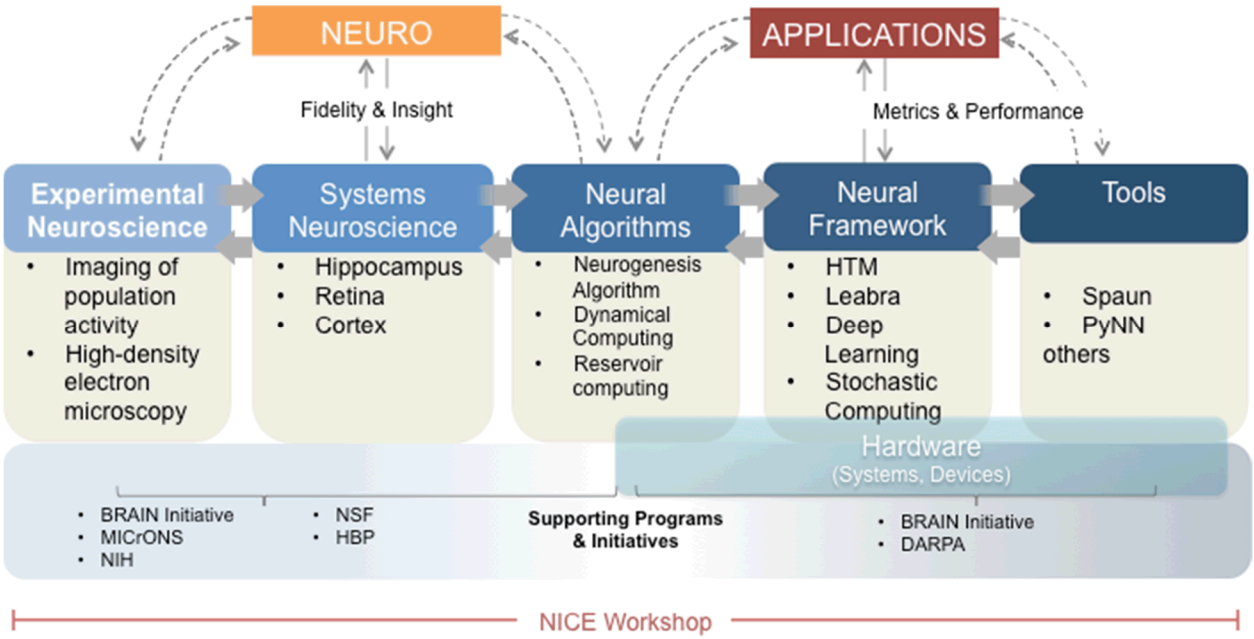

Figure 17: A wide spectrum of highly interdependent scientific and technical disciplines are at the core of neuro-inspired/neuromorphic computing approaches and applications, which were represented at the NICE Workshop. These disciplines interact at relatively local scales, and the NICE workshop has the goal of reaching across and providing the critical, longer-range interactions.

Application Areas

In this section, we identify and discuss the application areas for neuro-inspired computing. That "killer app" that drives technology development has yet to be identified, but it is likely to emerge from one of these application areas.

High Performance Computing with Hybrid Architectures

One of the critical application areas for new computing architectures is in High Performance Computing (HPC) where extremely large data sets and increasingly harder problems are being envisioned for the next generation of computing systems (i.e., Exascale). An interesting concept, which has been discussed at this workshop and other new computing technology-related venues, is a hybrid approach where the neuro-inspired systems can perform the functions that are too costly (in terms of energy, time and hardware) in conjunction with conventional computing systems that are specifically designed for numerical computation tasks. Such accelerators could be embedded in an HPC system to support the discovery and analysis tasks at higher energy-efficiency levels than a conventional system would be able to achieve without neuro-inspired algorithms.

Robotics and Control Systems

Next-generation microelectronic devices have enabled robotics (UAVs, industrial robots, self driving vehicles, etc.) to steadily increase their functionality and flexibility. Despite these improvements, the higher level functions are still performed by a human supervisor, usually necessitating a bi-directional high bandwidth data link between the platform and the supervisor or on-board personnel. While this can be accommodated in certain situations, availability of a higher functioning, self-contained, semi-autonomous system would greatly change the utilization scenarios in many use cases. Energy efficiency, real-time functionality and physical footprint of the systems are of critical concern on most applications where power, space, and response time are at a premium.

Big Data, High Velocity Data, Internet-of-Things

In many of the current consumer-facing and enterprise applications, data volume and the rate of data processing have stretched conventional computing resources to the limits. Especially in applications where the "sense-making" of the data analysis is not clearly defined or available, there is an opportunity for new approaches that can detect patterns and anomalies and guide further analysis by other, conventional means. In some scientific experiments, the data volume has been pre-compressed due to sensor characteristics or pre-processing algorithms. These endeavors could also benefit from a new approach to collecting, analyzing and interpreting the physical mechanisms underlying the observations, rather than relying on pre-determined compression, filtering and analysis approaches.

Cybersecurity

Cybersecurity is an example of a critical problem with extremely high data rates and volumes. New computational approaches could have immediate, high impact in this domain. Many current network management and cybersecurity protocols are reactionary: after an intrusion or problem is detected and analyzed, a countermeasure is developed and deployed. A higher level of functionality in anomaly detection and the ability to correlate multiple, seemingly unrelated features without being specifically programmed to do so could enable new approaches for the management and maintenance of networks at varying scales of deployment.

Neural Architecture, Theory and Algorithm Exploration

As discussed at the workshop in great depth, the datasets and experimental findings in neuroscience are vast, and not always well linked. The ability to formulate, test, verify or disprove theories around neural circuits, computation and other scientific questions with the current and emerging neuro-inspired and neuromorphic platforms at large scales (time, populations, variations in parameters, etc.) would be a critical capability which does not currently exist. The possibility of using these results to formulate better brain-machine interfaces, neural prosthesis, and medical treatments is a worthy goal.

Technical Approaches

This section briefly describes technical approaches to demonstrating neuro-inspired computing principles. Software implementation of neuro-inspired and neuromorphic algorithms tends to be the shortest path to a demonstration, and results achieved through software demonstration inform hardware development.

Software Implementation (Conventional Hardware)

Many of the current approaches to neuro-inspired computing (convolutional networks – machine learning, Hierarchical Temporal Memory (HTM), Leabra, Spaun) are implemented in software, on conventional computing resources. This implementation provides the highest level of flexibility and accessibility for the evaluation and further development of current approaches. The time required to perform the simulations or learning cycles, however, tend to be 10-100x slower than real time. Out of this category of work, great results are informing the following technical approaches.

Combining Software with Improved Architecture (New Devices)

Systems specifically designed and constructed to simulate and study neural circuits (SpiNNaker, BrainScaleS, Neurogrid, Field-Programmable Analog Arrays, IFAT, etc.) have been under development for around a decade, or longer. These systems provide significant improvements in certain metrics (energy efficiency, time to execute/model, biological realism) but give up some of the flexibility and familiarity that is available in the software based approaches. Workshop participants acknowledged that these systems are poised for large scale simulations and evaluation of theories, frameworks and algorithms generated by the community. These systems will also help to accelerate the discovery and development of the next generation of computing, control and analysis systems, and be the testing ground for new directions in neuroscience.

Novel Architectures

It might be possible to develop in hardware a completely new way of representing, processing, storing and recalling information, based on the sparse, hierarchical features of spiking neural networks. These systems are further away from the conventional, symbolic computation model but could be the key to breaking through some conventional computing barriers. Other technical approaches will critically inform how such systems can be built, optimized and utilized in a wide variety of applications. One distinct feature

is the absence of a clear "hardware/software" division. While there will still be digital, analog, electronic, optical and/or other novel devices that can be locally programmed and inspected (peek and poke), operation and functionality of the complete system will require new tools, metrics and interfaces.

Pathways and Resources

With Pathways and Resources, we identify at a high level how work is currently organized and funded.

Large Scale Programs (BRAIN Initiative, EU HBP)

Several existing large, multi-national programs are supporting a cross-cutting assembly of scientific fields including neuroscience and neuromorphic computing. Applications are generally viewed as a later-stage product of this activity, with scientific output being the primary goal. The interactions at the NICE workshop fostered new connections and strengthened existing collaborations. A sense of new, unexplored areas and great potential was tempered with a warning of previous epochs, "Third Wave of …" (Neural Networks, Neuromorphic Computing, New Computing Paradigm). Large scale efforts in the 1970s, '80s and following decades, which attempted to achieve higher levels of functionality than contemporary computers created the foundation for current efforts. A potential development path around high performance computing activities was also presented, which would further support activity in neuro-inspired and neuromorphic computing and provide specific application goals and metrics to drive development efforts.

New Research Projects at the Institutional Level

Many universities, research organizations and foundations have increased their levels of activity in neuroscience and related disciplines including neuro-inspired and neuromorphic computing, with some focusing on more of the early, scientific goals and others positioning themselves in various intersections of the wide spectrum of involved disciplines. There is a high level of interest in neuroscience and neuro- and bio-inspired themes in the academic realm, evidenced by numbers of both faculty and students who are getting involved and by expressions of interest by active researchers.

Commercial Development and Evaluation

In the commercial sector, organizations that routinely handle large volumes of data and derive value from processing, storing and analyzing these data sets, and providers of systems that enable this activity have been actively exploring alternative computing approaches. Machine learning and associated activity has been a great example of new functionality that was enabled by increasing computing power that in turn has provided improved methods to analyze data. No clear candidate for next-generation devices or architectures yet exists to supplement or replace conventional computing and CMOS microelectronics; early evaluation is valuable to both the systems manufacturers and service providers. We expect this "pre-competitive" arena will become even more

valuable and could see increasing support and activity related to commercial development.

Summary of 2015 NICE Workshop Findings

The final part of the workshop involved a breakout session at which participants discussed and helped articulate the most pertinent findings from the presentations and discussions. The results of this session shaped the following list of eight key findings:

Finding 1 – Neuro-inspired/neuromorphic systems could provide value in two, tightly coupled tracks: 1) as a new approach for analyzing, making sense of data, and predicting and controlling systems, 2) as a platform for understanding neural systems and testing hypotheses generated by neuroscience.

Finding 2 – Increasingly, the level of interest in brain-inspired computing approaches is moving beyond academic circles to broader government and industrial communities.

Finding 3 – Large scale projects and programs are underway in neuroscience, scientific computing, neural algorithm discovery, and hardware development and operation; and there is early interest in application metrics/definition for future systems evaluation.

Finding 4 – Although current machine learning and other neuro- or bio-inspired systems have demonstrated valuable functions, further developments are necessary to achieve the higher levels of functionality desired for wider-spectrum applications.

Finding 5 – High throughput techniques in experimental neuroscience are helping influence more advanced computational theories of neural function, but the community's capability to translate these computational neuroscience concepts into real mathematical theories and application-useful algorithms is still immature.

Finding 6 – Notable unresolved questions still facing the community include the level of neurobiological fidelity necessary for application impact, the necessity of future biological knowledge from experimentation to achieve neural computing's goals, and the best strategies for achieving learning (both theoretically and in real systems).

Finding 7 – The community appears to be approaching a general consensus that spike-based computation provides real, differentiating advantages over classic digital or analog approaches.

Finding 8 – There are several theories and frameworks that are ready for implementation (HTM, stochastic computing, Leabra, SPAUN, etc.) on the emerging neuromorphic and neuro-inspired computing platforms and they were presented at the workshop.

2015 NICE Workshop Recommendations

Relative to the key findings described in the previous section, and based on participant input during the workshop's breakout session, this report offers several recommendations. Some of these recommendations are intended to influence the organization of effort. Others are intended to influence the direction of research.

Recommendation 1 – Establish a coordinated effort.

For both conventional computing and emerging approaches, a coordinated effort across multiple disciplines and application areas is needed to: 1) establish appropriate metrics, 2) develop performance parameters with specific application cases to evaluate current systems, and 3) support future development of neuro-inspired/neuromorphic systems.

Recommendation 2 – Maintain the multi-disciplinary nature of NICE Workshops.

While seeing details of experimental neuroscience is not immediately useful to applications and seeing microelectronic device power consumption metrics are not critical for colleagues involved in biological experiments, having the visibility across the full spectrum is necessary and should be maintained in future NICE Workshops.

Recommendation 3 – Use neuro-inspired platforms to develop, test, and refine theories, algorithms and systems.

Current and emerging neuro-inspired platforms appear to have value today and should be used to help construct, test, verify and refine new computational neuroscience and machine learning theories. Likewise, there is value in developing novel hardware platforms, particularly those that incorporate online plasticity and low energy communication strategies.

Recommendation 4 – Quantify and communicate the value proposition.

The perceived value of applying existing neural algorithm frameworks to real-world applications should be quantified and communicated broadly. However, it is important to acknowledge that these are only the "tip of the iceberg" in neural algorithms, and the development of new neural algorithms from the community's growing knowledge of neuroscience is critical.

Recommendation 5 – Demonstrate specific application cases.

The community needs to develop a stronger application story. In lieu of the 'killer app,' which is not clearly visible or readily defined for neural-inspired systems today, the community should actively be developing clear application examples that demonstrate capabilities beyond current conventional computing approaches.

Recommendation 6 – Develop stronger mathematical neural theory.

Stronger mathematical neural theory is required. Improvements in this field will facilitate the transition of conceptual- or simulation-based computational neuroscience theories to application-useful machine learning tools.

9. Conclusions

In this report, we covered a range of process technologies and device architectures that have the potential to make revolutionary advances to the state of the art for computing. For process technologies, we recognize that pattern transfer capabilities at ever-decreasing feature size is the most important capability in a microelectronics fab. For that reason, we focused on EUV lithography, NIL, and DSA lithography as process technologies that can lead to tech surprise. Feature size reduction is one way to improve the speed of transistors. Using higher mobility materials is another, so we also covered channel engineering as a potentially game-changing capability.

For architectures, we covered new approaches to memory integration and optical interconnects. Both are poised to break out and be widely deployed in the next five years, with high impact. We also reported on the state of the art for neuro-inspired computing, which probably has a time horizon of closer to ten years to be used as a widely deployed technology but the potential impact is huge.

We covered each of these technologies not to be exhaustive, but to give the reader a sense for what is possible. In fact, in some cases, our own research roadmap is consistent with what we reported on here. Progress will continue to be swift and we encourage the reader to contact us for the latest information on these and related technologies.