Optics EXPRESS

Photonic switching in high performance datacenters [Invited]

QIXIANG CHENG,* SÉBASTIEN RUMLEY, MEISAM BAHADORI, AND KEREN BERGMAN

Lightwave Research Laboratory, Department of Electrical Engineering, Columbia University, New York, NY 10027, USA
**qc2228@columbia.edu*

Abstract: Photonic switches are increasingly considered for insertion in high performance datacenter architectures to meet the growing performance demands of interconnection networks. We provide an overview of photonic switching technologies and develop an evaluation methodology for assessing their potential impact on datacenter performance. We begin with a review of three categories of optical switches, namely, free-space switches, III-V integrated switches and silicon integrated switches. The state-of-the-art of MEMS, LCOS, SOA, MZI and MRR switching technologies are covered, together with insights on their performance limitations and scalability considerations. The performance metrics that are required for optical switches to truly emerge in datacenters are discussed and summarized, with special focus on the switching time, cost, power consumption, scalability and optical power penalty. Furthermore, the Pareto front of the switch metric space is analyzed. Finally, we propose a hybrid integrated switch fabric design using the III-V/Si wafer bonding technique and investigate its potential impact on realizing reduced cost and power penalty.

OCIS codes: (200.6715) Switching; (250.6715) Switching; (130.4815) Optical switching devices.

References and links

1. ANANDTECH, Available from: https://www.anandtech.com/show/11824/nvidia-ships-first-volta-dgx-systems.
2. N. A. Gawande, J. B. Landwehr, J. A. Daily, N. R. Tallent, A. Vishnu, and D. J. Kerbyson, “Scaling Deep Learning Workloads: NVIDIA DGX-1/Pascal and Intel Knights Landing,” in *2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)* (2017).
3. A. Putnam, A. Caulfield, E. Chung, D. Chiou, K. Constantinides, J. Demme, H. Esmaeilzadeh, J. Fowers, J. Gray, M. Haselman, S. Hauck, S. Heil, A. Hormati, J.-Y. Kim, S. Lanka, E. Peterson, A. Smith, J. Thong, P. Y. Xiao, D. Burger, J. Larus, G. P. Gopal, and S. Pope, “A Reconfigurable Fabric for Accelerating Large-Scale Datacenter Services,” IEEE Micro. **35**(3), 13–24 (2014).
4. N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, R. Boyle, P.-l. Cantin, C. Chao, C. Clark, J. Coriell, M. Daley, M. Dau, J. Dean, B. Gelb, T. Vazir Ghaemmaghami, R. Gottipati, W. Gulland, R. Hagmann, C. Richard Ho, D. Hogberg, J. Hu, R. Hundt, D. Hurt, J. Ibarz, A. Jaffey, A. Jaworski, A. Kaplan, H. Khaitan, D. Killebrew, A. Koch, N. Kumar, S. Lacy, J. Laudon, J. Law, D. Le, C. Leary, Z. Liu, K. Lucke, A. Lundin, G. MacKean, A. Maggiore, M. Mahony, K. Miller, R. Nagarajan, R. Narayanaswami, R. Ni, K. Nix, T. Norrie, M. Omernick, N. Penukonda, A. Phelps, J. Ross, M. Ross, A. Salek, E. Samadiani, C. Severn, G. Sizikov, M. Snelham, J. Souter, D. Steinberg, A. Swing, M. Tan, G. Thorson, B. Tian, H. Toma, E. Tuttle, V. Vasudevan, R. Walter, W. Wang, E. Wilcox, and D. H. Yoon, “In-Datacenter Performance Analysis of a Tensor Processing Unit,” in *Proceedings of the 44th Annual International Symposium on Computer Architecture* (2017), pp. 1–12.
5. D. Miller, “Device Requirements for Optical Interconnects to Silicon Chips,” Proc. IEEE **97**(7), 1166–1185 (2009).
6. G. Wang, D. G. Andersen, M. Kaminsky, K. Papagiannaki, T. S. E. Ng, M. Kozuch, and M. Ryan, “c-Through: part-time optics in data centers,” Comput. Commun. Rev. **41**(4), 327–338 (2010).
7. N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat, “Helios: a hybrid electrical/optical switch architecture for modular data centers,” Comput. Commun. Rev. **40**(4), 339–350 (2010).
8. N. Farrington, A. Forencich, P.-C. Sun, S. Fainman, J. Ford, A. Vahdat, G. Porter, and G. Papen, “A 10 us Hybrid Optical-Circuit/Electrical-Packet Network for Datacenters,” in *Optical Fiber Communication Conference/National Fiber Optic Engineers Conference 2013* (OSA, 2013).

#321055 https://doi.org/10.1364/OE.26.016022
Journal © 2018 Received 30 Jan 2018; revised 28 Mar 2018; accepted 12 Apr 2018; published 8 Jun 2018

9. M. Ghobadi, R. Mahajan, A. Phanishayee, N. Devanur, J. Kulkarni, G. Ranade, P.-A. Blanche, H. Rastegarfar, M. Glick, and D. Kilpe, "ProjecToR: Agile Reconfigurable Data Center Interconnect," in *Proceedings of the 2016 ACM SIGCOMM Conference* (2016), pp. 216–229.
10. A. M. Adel, A. S. P. K. Saleh, J. E. Bowers, and R. C. Alferness, "Elastic WDM Switching for Scalable Data Center and HPC Interconnect Networks," in *2016 21st OptoElectronics and Communications Conference (OECC)* (2016), pp. 1–3.
11. W. Miao, J. Luo, S. Di Lucente, H. Dorren, and N. Calabretta, "Novel flat datacenter network architecture based on scalable and flow-controlled optical switch system," Opt. Express **22**(3), 2465–2472 (2014).
12. A. Wonfor, H. Wang, R. V. Penty, and I. H. White, "Large Port Count High-Speed Optical Switch Fabric for Use Within Datacenters," J. Opt. Commun. Netw. **3**(8), A32 (2011).
13. X. Zhou, H. Liu, and R. Urata, "Datacenter optics: requirements, technologies, and trends (Invited Paper)," Chin. Opt. Lett. **15**(5), 120008 (2017).
14. K. Wen, P. Samadi, S. Rumley, C. P. Chen, Y. Shen, M. Bahadori, K. Bergman, and J. Wilke, "Flexfly: Enabling a Reconfigurable Dragonfly through Silicon Photonics," in *SC16: International Conference for High Performance Computing, Networking, Storage and Analysis* (2016).
15. S. Rumley, M. Bahadori, R. Polster, S. D. Hammond, D. M. Calhoun, K. Wen, A. Rodrigues, and K. Bergman, "Optical interconnects for extreme scale computing systems," Parallel Comput. **64**, 65–80 (2017).
16. C. Marxer and N. F. de Rooij, "Micro-opto-mechanical 2 × 2 switch for single-mode fibers based on plasma-etched silicon mirror and electrostatic actuation," J. Lightwave Technol. **17**(1), 2–6 (1999).
17. P. De Dobbelaere, K. Falta, S. Gloeckner, and S. Patra, "Digital MEMS for optical switching," IEEE Commun. Mag. **40**(3), 88–95 (2002).
18. M. Yano, F. Yamagishi, and T. Tsuda, "Optical MEMS for photonic switching-compact and stable optical crossconnect switches for simple, fast, and flexible wavelength applications in recent photonic networks," IEEE J. Sel. Top. Quantum Electron. **11**(2), 383–394 (2005).
19. L. Fan, S. Gloeckner, P. D. Dobblelaere, S. Patra, D. Reiley, C. King, T. Yeh, J. Gritters, S. Gutierrez, Y. Loke, M. Harburn, R. Chen, E. Kruglick, A. Husain, and M. Wu, "Digital MEMS switch for planar photonic crossconnects," in *Optical Fiber Communication Conference and Exhibit* (2002), pp. 93–94.
20. M. C. Wu and P. R. Patterson, *Free-space Optical MEMS*, E. J. G. Korvink and O. Paul, eds. (2005). p. 345–402.
21. R. Ryf, J. Kim, J. P. Hickey, A. Gnauck, D. Carr, F. Pardo, C. Bolle, R. Frahm, N. Basavanhally, C. Yoh, D. Ramsey, R. Boie, R. George, J. Kraus, C. Lichtenwalner, R. Papazian, J. Gates, H. R. Shea, A. Gasparyan, V. Muratov, J. E. Griffith, J. A. Prybyla, S. Goyal, C. D. White, M. T. Lin, R. Ruel, C. Mijander, S. Arney, D. T. Neilson, D. J. Bishop, P. Kolodner, S. Pau, C. Nuzman, A. Weis, B. Kumar, D. Lieuwen, V. Aksyuk, D. S. Greywall, T. C. Lee, H. T. Soh, W. M. Mansfield, S. Jin, W. Y. Lai, H. A. Huggins, D. L. Barr, R. A. Cirelli, G. R. Bogart, K. Teffeau, R. Vella, H. Mavoori, A. Ramirez, N. A. Ciampa, F. P. Klemens, M. D. Morris, T. Boone, J. Q. Liu, J. M. Rosamilia, and C. R. Gile, "1296-port MEMS transparent optical crossconnect with 2.07 petabit per second switch capacity," in *Optical Fiber Communication Conference and Exhibit* (2001), paper PD28.
22. M. C. Wu, O. Solgaard, and J. E. Ford, "Optical MEMS for Lightwave Communication," J. Lightwave Technol. **24**(12), 4433–4454 (2006).
23. R. R. A. Syms, "Scaling laws for MEMS mirror-rotation optical cross connect switch," J. Lightwave Technol. **20**(7), 1084–1094 (2002).
24. J. Kim, C. J. Nuzman, B. Kumar, D. F. Lieuwen, J. S. Kraus, A. Weiss, C. P. Lichtenwalner, A. R. Papazian, R. E. Frahm, N. R. Basavanhally, D. A. Ramsey, V. A. Aksyuk, F. Pardo, M. E. Simon, V. Lifton, H. B. Chan, M. Haueis, A. Gasparyan, H. R. Shea, S. Arney, C. A. Bolle, P. R. Kolodner, R. Ryf, D. T. Neilson, and J. V. Gates, "1100 x 1100 port MEMS-based optical crossconnect with 4-dB maximum loss," IEEE Photonics Technol. Lett. **15**(11), 1537–1539 (2003).
25. D. T. Neilson, R. Frahm, P. Kolodner, C. A. Bolle, R. Ryf, J. Kim, A. R. Papazian, C. J. Nuzman, A. Gasparyan, N. R. Basavanhally, V. A. Aksyuk, and J. V. Gates, "256 × 256 Port Optical Cross-Connect Subsystem," J. Lightwave Technol. **22**(6), 1499–1509 (2004).
26. J. E. Ford, V. A. Aksyuk, D. J. Bishop, and J. A. Walker, "Wavelength add-drop switching using tilting micromirrors," J. Lightwave Technol. **17**(5), 904–911 (1999).
27. Calient S Series Optical Circuit Switch, Available from: http://www.calient.net/products/s-series-photonic-switch/.
28. D. M. Marom, D. T. Neilson, D. S. Greywall, N. R. Chien-Shing Pai, V. A. Basavanhally, D. O. Aksyuk, F. Lopez, M. E. Pardo, Y. Simon, P. Low, P. Kolodner, and C. A. Bolle, "Wavelength-selective 1 × K switches using free-space optics and MEMS micromirrors: theory, design, and implementation," J. Lightwave Technol. **23**(4), 1620–1630 (2005).
29. J. Tsai and M. C. Wu, "A high port-count wavelength-selective switch using a large scan-angle, high fill-factor, two-axis MEMS scanner array," IEEE Photonics Technol. Lett. **18**(13), 1439–1441 (2006).
30. Z. Zhang, Z. You, and D. Chu, "Fundamentals of phase-only liquid crystal on silicon (LCOS) devices," Light Sci. Appl. **3**(10), e213 (2014).
31. M. Wang, L. Zong, L. Mao, A. Marquez, Y. Ye, H. Zhao, and F. Vaquero Caballero, "LCoS SLM Study and Its Application in Wavelength Selective Switch," Photonics **4**(2), 22 (2017).

32. S. Frisken, G. Baxter, D. Abakoumov, H. Zhou, I. Clarke, and S. Poole, “Flexible and grid-less wavelength selective switch using LCOS technology,” in *2011 Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference* (2011).
33. B. Robertson, H. Yang, M. M. Redmond, N. Collings, J. R. Moore, J. Liu, A. M. Jeziorska-Chapman, M. Pivnenko, S. Lee, A. Wonfor, I. H. White, W. A. Crossland, and D. P. Chu, “Demonstration of Multi-Casting in a 1 × 9 LCOS Wavelength Selective Switch,” J. Lightwave Technol. **32**(3), 402–410 (2014).
34. K. A. Williams, E. A. J. M. Bente, D. Heiss, Y. Jiao, K. Ławniczuk, X. J. M. Leijtens, J. J. G. M. van der Tol, and M. K. Smit, “InP photonic circuits using generic integration,” Photon. Res. **3**(5), B60 (2015).
35. S. Tanaka, S.-H. Jeong, S. Yamazaki, A. Uetake, S. Tomabechi, M. Ekawa, and K. Morito, “Monolithically Integrated 8:1 SOA Gate Switch With Large Extinction Ratio and Wide Input Power Dynamic Range,” IEEE J. Quantum Electron. **45**(9), 1155–1162 (2009).
36. I. White, E. T. Aw, K. Williams, H. Wang, A. Wonfor, and R. Penty, “Scalable optical switches for computing applications,” J. Opt. Netw. **8**(2), 215–224 (2009).
37. R. Stabile, A. Rohit, and K. A. Williams, “Monolithically Integrated 8 × 8 Space and Wavelength Selective Cross-Connect,” J. Lightwave Technol. **32**(2), 201–207 (2014).
38. R. Stabile, A. Albores-Mejia, and K. A. Williams, “Monolithic active-passive 16 × 16 optoelectronic switch,” Opt. Lett. **37**(22), 4666–4668 (2012).
39. Q. Cheng, A. Wonfor, J. Wei, R. V. Penty, and I. White, “Low-energy, high-performance lossless 8 × 8 SOA switch,” in Optical Fiber Communication Conference, OSA Technical Digest (online) (Optical Society of America, 2015), paper Th4E.6.
40. Q. Cheng, M. Ding, A. Wonfor, J. Wei, R. V. Penty, and I. H. White, “The Feasibility of Building a 64x64 Port Count SOA-Based Optical Switch,” in *2015 International Conference on Photonics in Switching (PS)* (2015), pp. 199–201.
41. A. Rohit, K. A. Williams, X. J. M. Leijtens, T. de Vries, Y. S. Oei, M. J. R. Heck, L. M. Augustin, R. Notzel, D. J. Robbins, and M. K. Smit, “Monolithic Multiband Nanosecond Programmable Wavelength Router,” IEEE Photonics J. **2**(1), 29–35 (2010).
42. A. Rohit, J. Bolk, X. J. M. Leijtens, and K. A. Williams, “Monolithic Nanosecond-Reconfigurable 4x4 Space and Wavelength Selective Cross-Connect,” J. Lightwave Technol. **30**(17), 2913–2921 (2012).
43. Q. Cheng, P. S. A. Rohit, A. Wonfor, R. V. Penty, I. White, and K. Williams, “Fast Dynamic Wavelength and Path Scheduling in a Monolithic 8 × 8 Switch,” in Optical Fiber Communication Conference. OSA Technical Digest (online) (Optical Society of America, 2014), paper M3E.2.
44. Q. Cheng, R. Stabile, A. Rohit, A. Wonfor, R. V. Penty, I. H. White, and K. A. Williams, “First Demonstration of Automated Control and Assessment of a Dynamically Reconfigured Monolithic 8 × 8 Wavelength-and-Space Switch,” J. Opt. Commun. Netw. **7**(3), A388 (2015).
45. K. Hamamoto, T. Anan, K. Komatsu, M. Sugimoto, and I. Mito, “First 8x8 semiconductor optical matrix switches using GaAs AlGaAs electro-optic guided-wave directional couplers,” Electron. Lett. **28**(5), 441 (1992).
46. R. Stabile, P. DasMahapatra, and K. A. Williams, “4 × 4 InP Switch Matrix With Electro-Optically Actuated Higher Order Micro-Ring Resonators,” IEEE Photonics Technol. Lett. **28**(24), 2874–2877 (2016).
47. S. K. Hiroki Kouketsu, N. Koyama, A. Takei, T. Taniguchi, Y. Matsushima, and K. Utaka, “High-speed and compact non-blocking 8 × 8 InAlGaAs InAlAs Mach-Zehnder-type optical switch fabric,” in *Optical Fiber Communication Conference 2014*, OSA Technical Digest (online) (Optical Society of America, 2014), paper M2K.3.
48. Y. Ueda, S. Nakamura, S. Fujimoto, H. Yamada, K. Utaka, T. Shiota, and T. Kitatani, “Polarization-Independent Low-Crosstalk Operation of InAlGaAs–InAlAs Mach–Zehnder Interferometer-Type Photonic Switch With Hybrid Waveguide Structure,” IEEE Photonics Technol. Lett. **21**(16), 1118–1120 (2009).
49. Q. Cheng, A. Wonfor, R. V. Penty, and I. H. White, “Scalable, Low-Energy Hybrid Photonic Space Switch,” J. Lightwave Technol. **31**(18), 3077–3084 (2013).
50. J. E. Zucker, K. L. Jones, T. H. Chiu, B. Tell, and K. Brown-Goebeler, “Strained quantum wells for polarization-independent electrooptic waveguide,” J. Lightwave Technol. **10**(12), 1926–1930 (1992).
51. Q. Cheng, A. Wonfor, J. L. Wei, R. V. Penty, and I. H. White, “Monolithic MZI-SOA hybrid switch for low-power and low-penalty operation,” Opt. Lett. **39**(6), 1449–1452 (2014).
52. M. Ding, A. Wonfor, Q. Cheng, R. V. Penty, and I. H. White, “Hybrid MZI-SOA InGaAs/InP Photonic Integrated Switches,” IEEE J. Sel. Top. Quantum Electron. **24**(1), 1–8 (2018).
53. S. Liu, Q. Cheng, M. R. Madarbux, A. Wonfor, R. V. Penty, I. H. White, and P. M. Watts, “Low Latency Optical Switch for High Performance Computing With Minimized Processor Energy Load,” J. Opt. Commun. Netw. **7**(3), A498 (2015).
54. Q. Cheng, A. Wonfor, J. L. Wei, R. V. Penty, and I. H. White, “Demonstration of the feasibility of large-port-count optical switching using a hybrid Mach-Zehnder interferometer-semiconductor optical amplifier switch module in a recirculating loop,” Opt. Lett. **39**(18), 5244–5247 (2014).
55. M. Bahadori, A. Gazman, N. Janosik, S. Rumley, Z. Zhu, R. Polster, Q. Cheng, and K. Bergman, “Thermal Rectification of Integrated Microheaters for Microring Resonators in Silicon Photonics Platform,” J. Lightwave Technol. **36**(3), 773–788 (2017).
56. K. Tanizawa, K. Suzuki, M. Toyama, M. Ohtsuka, N. Yokoyama, K. Matsumaro, M. Seki, K. Koshino, T. Sugaya, S. Suda, G. Cong, T. Kimura, K. Ikeda, S. Namiki, and H. Kawashima, “Ultra-compact 32 × 32 strictly-non-blocking Si-wire optical switch with fan-out LGA interposer,” Opt. Express **23**(13), 17599–17606 (2015).

57. L. Qiao, W. Tang, and T. Chu, “32 × 32 silicon electro-optic switch with built-in monitors and balanced-status units,” Sci. Rep. **7**, 42306 (2017).
58. D. Celo, D. J. Goodwill, J. Jiang, P. Dumais, C. Zhang, F. Zhao, X. Tu, C. Zhang, S. Yan, J. He, M. Li, W. Liu, Y. Wei, D. Geng, H. Mehrvar, and E. Bernier, "32×32 silicon photonic switch," in *21st OptoElectronics and Communications Conference (OECC)* (2016), pp. 1–3.
59. L. Lu, S. Zhao, L. Zhou, D. Li, Z. Li, M. Wang, X. Li, and J. Chen, “16 × 16 non-blocking silicon optical switch based on electro-optic Mach-Zehnder interferometers,” Opt. Express **24**(9), 9295–9307 (2016).
60. L. Qiao, T. Wang, and T. Chu, “16 × 16 Non-blocking Silicon Electro-optic Switch Based on Mach-Zehnder Interferometers,” in *Optical Fiber Communication Conference, OSA Technical Digest* (online) (Optical Society of America, 2016), paper Th1C.2.
61. Y. Huang, Q. Cheng, and K. Bergman, “Automated Calibration of Balanced Control to Optimize Performance of Silicon Photonic Switch Fabrics,” in *Optical Fiber Communication Conference*, OSA Technical Digest (online) (Optical Society of America, 2018). paper Th1G.2.
62. N. Dupuis, B. G. Lee, A. V. Rylyakov, D. M. Kuchta, C. W. Baks, J. S. Orcutt, D. M. Gill, W. M. J. Green, and C. L. Schow, “Design and Fabrication of Low-Insertion-Loss and Low-Crosstalk Broadband 2x2 Mach-Zehnder Silicon Photonic Switches,” J. Lightwave Technol. **33**(17), 3597–3606 (2015).
63. N. Dupuis, A. V. Rylyakov, C. L. Schow, D. M. Kuchta, C. W. Baks, J. S. Orcutt, D. M. Gill, W. M. Green, and B. G. Lee, “Ultralow crosstalk nanosecond-scale nested 2 × 2 Mach-Zehnder silicon photonic switch,” Opt. Lett. **41**(13), 3002–3005 (2016).
64. Z. Lu, D. Celo, H. Mehrvar, E. Bernier, and L. Chrostowski, “High-performance silicon photonic tri-state switch based on balanced nested Mach-Zehnder interferometer,” Sci. Rep. **7**(1), 12244 (2017).
65. L. Schares, T. N. Huynh, M. G. Wood, R. Budd, F. Doany, D. Kuchta, N. Dupuis, B. G. Lee, C. L. Schow, M. Moehrle, A. Sigmund, W. Rehbein, T. Y. Liow, L. W. Luo, and G. Q. Lo, "A Gain-Integrated Silicon Photonic Carrier with SOA-Array for Scalable Optical Switch Fabrics," in *Optical Fiber Communication Conference*, OSA Technical Digest (online) (Optical Society of America, 2016), paper Th3F.5.
66. R. Konoike, K. Suzuki, T. Inoue, T. Matsumoto, T. Kurahashi, A. Uetake, K. Takabayashi, S. Akiyama, S. Sekiguchi, K. Ikeda, S. Namiki, and H. Kawashima, “Lossless Operation of SOA-Integrated Silicon Photonics Switch for 8 × 32-Gbaud 16-QAM WDM Signals,” in *Optical Fiber Communication Conference Postdeadline Papers* (OSA, 2018).
67. Q. Cheng, M. Bahadori, Y. Huang, S. Rumley, and K. Bergman, “Smart Routing Tables for Integrated Photonic Switch Fabrics,” in *2017 European Conference on Optical Communication* (ECOC) (2017), paper M.1.A.2.
68. Q. Cheng, M. Bahadori, and K. Bergman, "Advanced Path Mapping for Silicon Photonic Switch Fabrics," in *Conference on Lasers and Electro-Optics*, OSA Technical Digest (online) (Optical Society of America, 2017), paper SW1O.5.
69. M. Bahadori, S. Rumley, H. Jayatilleka, K. Murray, N. A. F. Jaeger, L. Chrostowski, S. Shekhar, and K. Bergman, “Crosstalk Penalty in Microring-Based Silicon Photonic Interconnect Systems,” J. Lightwave Technol. **34**(17), 4043–4052 (2016).
70. Y. Huang, Q. Cheng, Nathan,C. Abrams, J. Zhou, S. Rumley, K. Bergman, Automated Calibration and Characterization for Scalable Integrated Optical Switch Fabrics without Built-in Power Monitors, in 2017 European Conference on Optical Communication (ECOC) (2017), paper M.1.A.3.
71. Q. Xu, B. Schmidt, S. Pradhan, and M. Lipson, “Micrometre-scale silicon electro-optic modulator,” Nature **435**(7040), 325–327 (2005).
72. N. Sherwood-Droz, H. Wang, L. Chen, B. G. Lee, A. Biberman, K. Bergman, and M. Lipson, “Optical 4x4 hitless slicon router for optical networks-on-chip (NoC),” Opt. Express **16**(20), 15915–15922 (2008).
73. B. G. Lee, A. Biberman, P. Dong, M. Lipson, and K. Bergman, “All-Optical Comb Switch for Multiwavelength Message Routing in Silicon Photonic Networks,” IEEE Photonics Technol. Lett. **20**(10), 767–769 (2008).
74. A. Biberman, H. L. R. Lira, K. Padmaraju, N. Ophir, J. Chan, M. Lipson, and K. Bergman, “Broadband Silicon Photonic Electrooptic Switch for Photonic Interconnection Networks,” IEEE Photonics Technol. Lett. **23**(8), 504–506 (2011).
75. P. DasMahapatra, R. Stabile, A. Rohit, and K. A. Williams, “Optical Crosspoint Matrix Using Broadband Resonant Switches,” IEEE J. Sel. Top. Quantum Electron. **20**(4), 1–10 (2014).
76. D. Nikolova, D. M. Calhoun, Y. Liu, S. Rumley, A. Novack, T. Baehr-Jones, M. Hochberg, and K. Bergman, “Modular architecture for fully non-blocking silicon photonic switch fabric,” Microsystems & Nanoengineering **3**, 16071 (2017).
77. A. S. P. Khope, T. Hirokawa, A. M. Netherton, M. Saeidi, Y. Xia, N. Volet, C. Schow, R. Helkey, L. Theogarajan, A. A. M. Saleh, J. E. Bowers, and R. C. Alferness, “On-chip wavelength locking for photonic switches,” Opt. Lett. **42**(23), 4934–4937 (2017).
78. K. Padmaraju, D. F. Logan, T. Shiraishi, J. J. Ackert, A. P. Knights, and K. Bergman, “Wavelength Locking and Thermally Stabilizing Microring Resonators Using Dithering Signals,” J. Lightwave Technol. **32**(3), 505–512 (2014).
79. X. Zhu, K. Padmaraju, L.-W. Luo, S. Yang, M. Glick, R. Dutt, M. Lipson, and K. Bergman, “Fast Wavelength Locking of a Microring Resonator,” IEEE Photonics Technol. Lett. **26**(23), 2365–2368 (2014).
80. Y. Vlasov, W. M. J. Green, and F. Xia, “High-throughput silicon nanophotonic wavelength-insensitive switch for on-chip optical networks,” Nat. Photonics **2**(4), 242–246 (2008).

81. Q. Cheng, M. Bahadori, S. Rumley, and K. Bergman, “Highly-scalable, low-crosstalk architecture for ring-based optical space switch fabrics,” in *2017 IEEE Optical Interconnects Conference* (OI, 2016), p. 41–42.
82. T. J. Seok, N. Quack, S. Han, R. S. Muller, and M. C. Wu, “Highly Scalable Digital Silicon Photonic MEMS Switches,” J. Lightwave Technol. **34**(2), 365–371 (2016).
83. T. J. Seok, N. Q. S. Han, W. Zhang, R. S. Muller, M. C. Wu, “64x64 Low-Loss and Broadband Digital Silicon Photonic MEMS Switches,” in *2015 European Conference on Optical Communication (ECOC)* (2015).
84. Y. Shen, M. H. N. Hattink, P. Samadi, Q. Cheng, Z. Hu, A. Gazman, and K. Bergman, "Software-defined networking control plane for seamless integration of multiple silicon photonic switches in Datacom networks," Opt. Express **26**, 10914-10929 (2018).
85. M. Y. Teh, J. J. Wilke, K. Bergman, and S. Rumley, *Design Space Exploration of the Dragonfly Topology* (Springer International Publishing, 2017).
86. K. Suzuki, R. Konoike, J. Hasegawa, S. Suda, H. Matsuura, K. Ikeda, S. Namiki, and H. Kawashima, “Low Insertion Loss and Power Efficient 32 × 32 Silicon Photonics Switch with Extremely-High-Δ PLC Connector,” in *Optical Fiber Communication Conference Postdeadline Papers* (2018).
87. T. Barwicz, Y. Taira, T. W. Lichoulas, N. Boyer, Y. Martin, H. Numata, J.-W. Nah, S. Takenobu, A. Janta-Polczynski, E. L. Kimbrell, R. Leidy, M. H. Khater, S. Kamlapurkar, S. Engelmann, Y. A. Vlasov, and P. Fortier, “A Novel Approach to Photonic Packaging Leveraging Existing High-Throughput Microelectronic Facilities,” IEEE J. Sel. Top. Quantum Electron. **22**(6), 455–466 (2016).
88. M. R. T. Tan, P. Rosenberg, W. V. Sorin, B. Wang, S. Mathai, G. Panotopoulos, and G. Rankin, “Universal Photonic Interconnect for Data Centers,” J. Lightwave Technol. **36**(2), 175–180 (2018).
89. M. J. R. Heck, J. F. Bauters, M. L. Davenport, J. K. Doylend, S. Jain, G. Kurczveil, S. Srinivasan, Y. Tang, and J. E. Bowers, “Hybrid Silicon Photonic Integrated Circuit Technology,” IEEE J. Sel. Top. Quantum Electron. **19**(4), 6100117 (2013).
90. S. Stankovic, R. Jones, M. N. Sysak, J. M. Heck, G. Roelkens, and D. Van Thourhout, “Hybrid III–V/Si Distributed-Feedback Laser Based on Adhesive Bonding,” IEEE Photonics Technol. Lett. **24**(23), 2155–2158 (2012).
91. L. Chen, E. Hall, L. Theogarajan, and J. Bowers, “Photonic Switching for Data Center Applications,” IEEE Photonics J. **3**(5), 834–844 (2011).
92. L. Schares, R. Budd, D. Kuchta, F. Doany, C. Schow, M. Möhrle, A. Sigmund, and W. Rehbein, “Etched-facet semiconductor optical amplifiers for gain-integrated photonic switch fabrics,” in *2015 European Conference on Optical Communication (ECOC)* (OSA, 2015).
93. H. Park, A. W. Fang, O. Cohen, R. Jones, M. J. Paniccia, and J. E. Bowers, “A Hybrid AlGaInAs–Silicon Evanescent Amplifier,” IEEE Photonics Technol. Lett. **19**(4), 230–232 (2007).
94. M. L. Davenport, S. Skendzic, N. Volet, and J. E. Bowers, “Heterogeneous Silicon/InP Semiconductor Optical Amplifiers with High Gain and High Saturation Power,” in *Conference on Lasers and Electro-Optics* (OSA, 2016).
95. B. Ben Bakir, A. Descos, N. Olivier, D. Bordel, P. Grosse, E. Augendre, L. Fulbert, and J. M. Fedeli, “Electrically driven hybrid Si/III-V Fabry-Pérot lasers based on adiabatic mode transformers,” Opt. Express **19**(11), 10317–10325 (2011).
96. S. Keyvaninia, S. Verstuyft, L. Van Landschoot, F. Lelarge, G. H. Duan, S. Messaoudene, J. M. Fedeli, T. De Vries, B. Smalbrugge, E. J. Geluk, J. Bolk, M. Smit, G. Morthier, D. Van Thourhout, and G. Roelkens, “Heterogeneously integrated III-V/silicon distributed feedback lasers,” Opt. Lett. **38**(24), 5434–5437 (2013).
97. K. Tanizawa, K. Suzuki, S. Suda, H. Matsuura, K. Ikeda, S. Namiki, and H. Kawashima, “Silicon photonic 32 × 32 strictly-non-blocking blade switch and its full path characterization,” in *2016 21st OptoElectronics and Communications Conference (OECC)* (2016).
98. Q. Cheng, A. Wonfer, R. V. Penty, and I. H. White, “Robust Large-Port-Count Hybrid Switches with Relaxed Control Tolerances,” in *CLEO: 2015*, OSA Technical Digest (online) (Optical Society of America, 2015), paper JTh2A.38.
99. A. Fontcuberta i Morral, J. M. Zahler, H. A. Atwater, S. P. Ahrenkiel, and M. W. Wanlass, “InGaAs/InP double heterostructures on InP/Si templates fabricated by wafer bonding and hydrogen-induced exfoliation,” Appl. Phys. Lett. **83**(26), 5413–5415 (2003).

1. Introduction

Datacenter networks can be abstracted as three major functional blocks that include: (1) a collection of links for transmitting information across the numerous end-points; (2) network interfaces, responsible for injecting and collecting information to and from each link; and (3) switches, for gearing information received on the input links to the appropriate output links. The continued growth in raw performance of datacenter servers is driving increasing bandwidth demands on all three of these network functional blocks. For instance, the DGX Station from Nvidia, commercially introduced in 2017 [1], and optimized for machine learning operations in datacenters, is equipped with no less than four Infiniband (EDR 100 Gb/s) interfaces [2]. This quad-rail connectivity had to be selected as the next generation

Infiniband standard (HDR – 200 Gb/s) wasn't available at the time. Servers with yet higher bandwidth requirements can reasonably be expected in the next years, in particular as more specialized computing hardware is developed [3,4].

Electrical transmissions over several centimeters are roughly limited to tens of Gb/s on a single wire [5]. Thus, reaching hundreds of Gb/s of throughput requires a high level of parallelism which in turn leads to bulky interfaces and designs of increased complexity. Moreover, approaching the end of Moore's Law, the number of transistors available per unit area is saturating, and realizing electronic switches with a large number of ports each offering hundreds of Gb/s becomes increasingly challenging. To nevertheless support the ever growing bandwidth demands in datacenters, researchers have looked for novel link and interface architectures, and, central to this article, for alternative ways of performing switching, in particular using optical switches [6–12]. On the link level, for example, pluggable optical transceivers are already at 100 Gb/s and are targeted to provide 20 Tb/s of front panel optical bandwidth density for datacenters in 2020 with a scaling trend of × 10 every seven years. On the switch level, the optical switching bandwidth density is predicted to have a steeper scaling trend than optical links at × 10 every five years [13].

The implementation of optical switching in datacenters imposes new challenges on the architecture of the network that must be overcome. The technological absence of optical buffering results in packet drops in cases of traffic contention in the network. For this reason, it appears very unlikely that true optical packet switching (that is, switches that conserve packets in the optical form) will emerge in the near term. Optical switches thus cannot be considered as a one-to-one replacement for electronic packet switches. If they are to be used in datacenters, the network architecture will likely employ optical switches in combination with conventional electronic switches. Some prior examples include the c-through [6], helios [7], and mordia [8] architectures. In all these examples, optical switches are used to adapt the network topology to specific traffic patterns. In other words, pairs of servers or racks exchanging a lot of traffic can be awarded additional direct optical connections, and thus more bandwidth. The connections, however, must be "stolen" from server or rack pairs exchanging little or no traffic at all [14].

Several technologies can enable spatial or wavelength-selective optical switches for datacenter applications, including micro-electromechanical systems (MEMS), liquid crystals on silicon (LCOS), semiconductor optical amplifiers (SOA), Mach-Zehnder interferometers (MZI) and micro-ring resonators (MRR). The choice of the optical switch for a particular application in a datacenter is ultimately driven by metrics such as cost-per-port, reconfiguration time and optical power penalties [15]. Using optical switches in combination with regular electrical ones consists essentially of solving network bottlenecks by re-allocating existing bandwidth instead of provisioning additional bandwidth. Consequently, the cost associated with bandwidth reallocation (i.e. the cost of the optical switches for the most part) must not exceed the one of additional bandwidth. The reconfiguration speed of the switch is also important as it (partially, as discussed hereafter) dictates the rate at which re-allocations can occur. The power penalties inflicted may prevent the utilization of transceivers with tight optical budgets. In summary, not only a thorough understanding of all the optical technologies is required to evaluate each individual metric, but also a combined analysis of all these metrics in the datacenter system is necessary to converge on the right choice for the optical switch.

In this paper, we first perform a technical review of the current technologies for optical switches in datacenters. This includes free-space switches (section 2), III-V integrated switches (section 3), and silicon integrated switches (section 4). We then present a metrics analysis (section 5) of optical switches for datacenter applications, including switching time, optical power penalty, cost, scalability, and the power consumption. In section 6, we present a hybrid switch design using the III-V/Si wafer bonding technique and discuss its potential impact on reducing the cost and optical power penalty. Finally, a summary is presented in section 7 and conclusions are drawn.

2. Free-space optical switches

Free-space optical switches have been realized by a number of competing technologies, including MEMS and LCOS. Both have been commercialized.

2.1 MEMS-based optical switches

MEMS-based optical switches are the most common and mature free-space switching devices. The micro-mirrors can be fabricated by either bulk micro-machining where the mechanical structures are etched directly on the silicon substrate [16], or surface micro-machining, in which epitaxial layers, such as polysilicon, silicon nitride and silicon oxide, are deposited, patterned and selectively removed [17]. Electrostatic driver is most commonly applied because of its low power consumption and ease of control. The typical yielding voltage however is high, reaching up to 100-150 V [18].

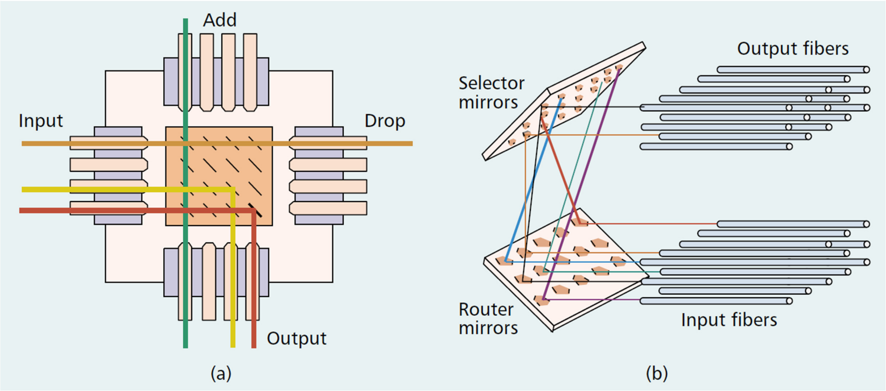

Fig. 1. (a) 2-D MEMS optical switch. (b) 3-D MEMS optical switch. Reprinted from [19].

The MEMS spatial switches were realized in both two-dimensional (2-D) and three-dimensional (3-D) configurations, as shown by Fig. 1a and 1b, respectively. The 2-D MEMS switches are implemented in the crossbar topology and operate digitally as the mirror position is bi-stable. Collimators are used, and to minimize the distribution of optical insertion loss due to the different optical path lengths, a confocal geometry is applied by equaling the average optical path length to the Rayleigh length [20]. An N × N 2-D MEMS switch requires N^2 micro-mirrors and its scalability is ultimately limited by the fabrication precision of the micro-mirrors, including their angle, size, fill factor, and curvature [20]. This is because larger mirrors are required for higher port count switches to support longer Rayleigh lengths [20]. OMM Inc. has commercialized up to 16 × 16 2-D MEMS switches [17]. Each individual switch cell had a micro-mirror assembled at 90 degree and sat on the actuator plate. The device exhibited an insertion loss of <6 dB with variations <l dB over the entire 1280-1650 nm wavelength range [19]. Crosstalk was measured to be <-50 dB and maximum switching time was 12 ms.

3-D MEMS switches were proposed to support very large-scale optical cross-connect devices [21]. This type of device is assembled by using 2-D input and output fiber arrays with collimators. Two stages of independent 2-D micro-mirror arrays are used to steer the optical beams in three dimensions and this requires micro-mirrors implemented with a two-axis tilting structure. The 3-D MEMS switch is advantageous in the scaling law of port count comparing to the 2-D devices: for a given mirror rotation angle, the mirror size scales at $\sqrt{N}$ whereas the 2-D mirror chip has a linear scale in dimension of N for an N × N switch system [22]. The ultimate port count in such a 3-D cross connect system is determined by the diameter of beams and the mirror rotation angle [23]. The crosstalk is inherently low which is caused by the small amount of light diffracted. Instead of simply increasing the overall size of the switch to achieve large port counts, improved optical layouts by inserting Fourier transform lens [24,25] and roof-type retroreflector [18] were proposed for compact installations and smaller mirror tilting angles. Lucent placed the Fourier transform lens with a focal length comparable to the Rayleigh length of the beam between the two micro-mirror array chips [24,25]. Up to 1100 × 1100 port count switch was demonstrated with a maximum insertion loss of 4.0 dB [24]. Switching time was reported at ~5 ms [25].

Switching at the wavelength granularity can be achieved in a 'disperse-and-select' architecture by using diffraction grating as free-space spectrometer [26,27]. Scalability of the MEMS-based WSSs depends on both the mirror tilting angle and the grating diffraction. 1 × 4 WSS with up to 128 channels spaced at 50 GHz was reported by Lucent with a fiber-to-fiber insertion loss of ~5 dB [28]. Further scaling on the port count can be done by arranging the output collimators in a 2D array, increasing the limit from N to N^2 [29].

2.2 LCOS-based optical switches

The light modulating properties of liquid crystal (LC) material has been explored for both amplitude modulation and phase modulation. The latter takes advantage of the birefringence of the liquid crystal material for phase manipulation [30]. The variation in optical properties, such as polarization and refractive index, can be adjusted by exerting voltages across the LC material to change the certain relative orientation of molecules. LCOS combines the light modulating feature of LC material and the advanced CMOS technology. Nematic LCOS devices are becoming the dominant technology for phase-only spatial light modulators with reconfiguration times in the range of 10-100 ms [31].

The phase-only LCOS device consists of a transparent top glass substrate, indium tin oxide electrode, alignment layers, LC material, aluminum pixel mirror array and CMOS backplane [30]. The incident light propagates through the LC with almost no absorption and is then reflected by the aluminum pixel mirrors for binary-phase hologram to increase diffraction efficiency. Each pixel is connected to the electrical circuitry buried underneath to receive control signals for the required hologram patterns. The LCOS WSS is also based on the 'disperse-and-switch' architecture. The light is launched from one of the input fiber array, and then directed to the diffraction grating and angularly dispersed. The colored beams are subsequently fed to the corresponding pixels on the LCOS, which act as a phase-only diffraction grating with a tunable period and pattern. This generates a deflection angle for each of the incident beams at a wavelength to route to the chosen output fiber. Insertion loss mainly comes from the reflection and polarization modulation of LCOS while crosstalk is caused by higher orders of diffraction because of the imperfection of the hologram, and the finite spatial and phase quantization [31]. The unique feature of LCOS-based optical switches is the grid-less capability, and more importantly, activating this feature requires no additional cost and does not compromise the performance and reliability [32]. Robertson *et al.* has reported a multi-function 1 × 9 switching system LCOS WSS [33], in which the nematic LCOS consisted an array of 1280 × 720 pixels with a pitch of 15 μm. The device was able to resolve 256 discrete phase levels with a maximum modulation of 2.2 π at 1550 nm. Insertion loss and the worst-case crosstalk was 7.6 dB and −19.4 dB, respectively, for a channel spacing of 100 and 200 GHz. Increasing the port count requires a higher deflecting ability of LCOS [31].

3. III-V-based optical switches

InP-based generic integration has enabled on-chip systems with increasing complexity. High-performance active components, such as optical amplifiers and lasers, are the unique selling point [34]. Optical switching circuits in the InP-platform have primarily applied SOA gated and interferometric switching elements.

3.1 SOA-based optical switches

SOA-gate based integrated switch fabrics have been mainly implemented in the broadcast-and-select and wavelength-selective configurations. Each path can be gated by one SOA element and its inherent gain opportunely overcomes the fan-out/fan-in losses while its high ON/OFF extinction ratio ensures excellent crosstalk suppression. SOA-gated switches are usually designed for single polarization, i.e. TE, operation; however, by introducing tensile-strain in multi-quantum-wells (MQWs), polarization insensitive SOAs can be realized [35]. Broadcast-and-select topology scales with a square law increase of switch elements and a two-fold increase of 1 × 2 splitters and combiners per path, discouraging scaling beyond 4 × 4 connectivity for monolithic integration.

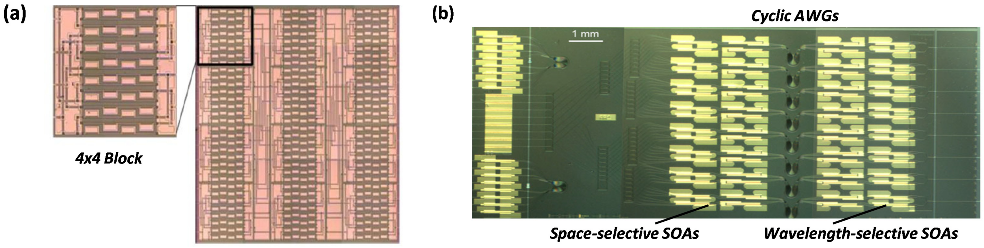

Fig. 2. (a) 16 × 16 InP-based all-active SOA switch. Reprinted from [12]. (b) 8 × 8 × 8λ InP-based wavelength-and-space SOA switch. Reprinted from [37].

Multistage architectures involving cascaded SOA elements certainly enable larger switch fabrics, but require additional examination of architecture design. A hybrid multistage architecture implementing broadcast-and-select sub-stages was proposed [36] to carefully establish a trade-off between the number of stages and the losses incurred within each switch stage. The first monolithically-integrated three-stage 16 × 16 SOA-based switch with 480 integrated components was subsequently demonstrated by the same group (Fig. 2a) [12]. This regrowth-free all-active device applied compact total internal reflection (TIR) mirrors and leveraged amplifying waveguides to provide lossless operation. Shortly after, an equivalent switch circuit with a passive-active integration by wafer regrowth was reported by Stabile *et al.* [38] for enhanced power efficiency and optical signal-to-noise-ratio (OSNR). However, improved performance and/or further scale-up would require a large reduction in component-level excess losses, a more careful design of balancing the summed loss with the SOA gain per stage, and a close examination of SOA designs for linear operation. A moderate-scale 8 × 8 SOA-gate switch was later fabricated in the same implementation demonstrating on-chip net gain and wide input power dynamic range (IPDR) [39], as well as showing the feasibility to construct a 64 × 64 SOA switch [40]. Monolithic integration of such a large device thus remains challenging due to the difficulty of a uniform wafer process.

WSS for SOA gated devices can be achieved by co-integrating arrayed waveguide gratings (AWGs) on-chip [41]. The cyclic AWGs act as on-chip de-multiplexers and each of their output ports follows an SOA gate to allow arbitrary combinations of wavelengths to be selected and routed. A monolithic implementation of such a wavelength-and-space switching circuit was proposed [42]. This concept consisted of a broadcast-and-select stage, a wavelength-selective stage and a broadband fan-in stage, and up to 8 × 8 × 8λ connectivity has been demonstrated (Fig. 2b) [37]. Dynamic routing with real-time path reallocation has been performed with microsecond time slots [43,44].

3.2 Interferometric optical switches

Interferometric switch fabrics eliminate the passive splitting/combining losses owing to their optically coherent operation in which their operating principles rely on the modulation of the refractive index. The response of interferometric switches is not digital, and thus a precise control of the bias conditions is required. Early demonstrations included the use of directional couplers (DCs) and MZIs. DC-based switches were integrated with up to 8 × 8 connectivity on GaAs substrate [45], but the manufacturing control of the coupling length was proved difficult. MRR element was also explored and a recent demonstration showed a 4 × 4 circuit [46]. MZIs represent the most mature switch technology with the geometrical separation of mode-coupling and phase-shifting regions, which allows separate optimization. Recent demonstration showed an 8 × 8 MZI switch fabric arranged in the N-stage planar topology (Fig. 3a) [47], representing the largest integrated MZI circuit in InP-platform so far. Polarization-independent operation can be achieved thanks to the plasma dispersion effect that provides a nearly similar modulation for TE and TM modes. Single 2 × 2 element exhibited crosstalk at the level of −20 dB for both polarizations [48], however, this number degraded to −11 dB for TE mode in the 8 × 8 device [47], which was likely due to the fabrication variability. The poor crosstalk brings about significant signal degradation that severely limits the scalability. Another limiting factor is the high insertion loss.

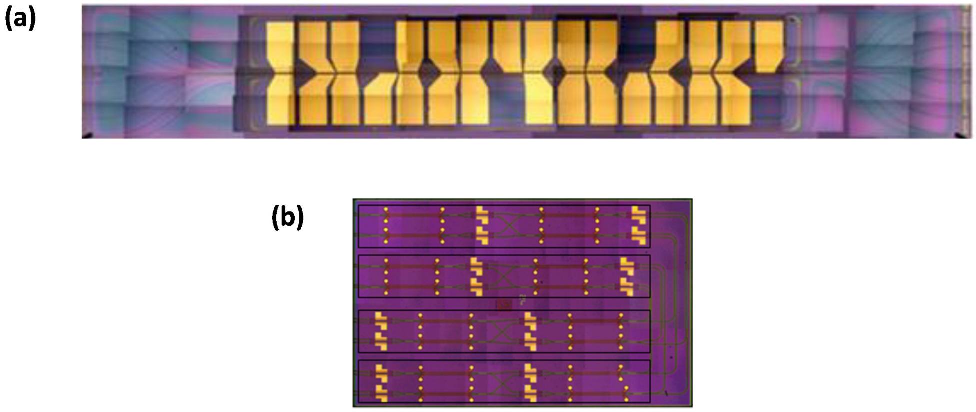

Fig. 3. (a) 8 × 8 InP-based MZI switch. Reprinted from [47]. (b) 4 × 4 InP-based hybrid MZI-SOA switch. Reprinted from [45].

To address the issues of crosstalk and insertion loss, an approach combining MZI switch elements with SOA gates was proposed [49]. Additional short SOA gates are introduced to the outputs of each MZI element in a dilated scheme for both crosstalk suppression and loss compensation (Fig. 3b). This has the further advantage that the short distributed SOAs only induce small signal impairment at each stage. Quantum confined Stark effect (QCSE)-based phase shifters were applied to facilitate low-energy switching through reverse bias. This often polarization sensitive effect can be modified to be polarization insensitive by using strained quantum-wells [50]. The integrated InP-based MZI-SOA switch fabrics exhibited excellent crosstalk ratio of better than −40 dB, negligible on-chip loss and ultra-wide IPDR with tiny power penalty floor [51,52]. A full analysis of power consumption of such switches used for computing applications can be found in [53]. This approach enabled the feasibility of building optical switch fabrics with up to 128 × 128 port counts [54]. As the length of the QCSE-based phase shifters often at the level of 1 mm [49], the ultimate size of such a large-scale switch will also be limited by the maximum available InP wafer size.

4. Silicon-based optical switches

The highly advanced CMOS industry with mature fabrication and manufacturing infrastructures has triggered a booming development in silicon photonics on silicon-on-isolator (SOI) platform. The large index contrast between the core (silicon) and cladding (silica) layers enables a strong confinement of the lightwave and thus leads to a much smaller footprint than on the InP platform. Silicon exhibits a strong thermo-optic (T-O) coefficient

(1.8×10^{-4} K^{-1}), and we have showed that this can be leveraged to tune the phase in tens of microseconds [55]. However, silicon does not possess linear electro-optic effects and its quadratic effects are very weak. To benefit from nanosecond-scale switching times, the plasma dispersion effect through carrier injection or depletion offers the best all-silicon solution for electro-optic (E-O) switch fabrics. Silicon-based optical switches have been mainly explored with interferometric and resonant switching elements. A new type of silicon integrated MEMS switch was recently developed.

4.1 MZI-based optical switches

Over the last few years silicon photonic integration technology has quickly matured to the point that up to tens of thousands of components can be monolithically integrated to realize increasingly sophisticated on-chip functionalities. The first monolithic 32 × 32 silicon-based MZI switch fabric was reported in 2015 (Fig. 4a) [56]. It implemented 1024 MZI components using T-O phase tuners and the footprint was reduced 45 times compared to that of the silica-based switching circuit. This device showed an excellent uniformity in on-chip path losses of 15.8 ± 1 dB and crosstalk ratios of <-35 dB benefitting from the usage of path-independent loss (PILOSS) topology. A response time of less than 30 µs was obtained. One year later, another 32 × 32 T-O MZI switch fabric co-integrating ~900 photodiodes for real-time calibration on each switch cell was demonstrated [58]. The device applied a custom dilated non-blocking topology and exhibited on-chip path losses of >23 dB, with rise/fall times at approximately 750 µs.

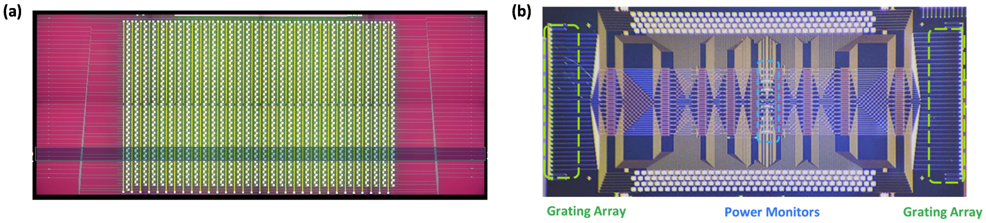

Fig. 4. (a) 32 × 32 silicon T-O MZI-based PILOSS switch. Reprinted from [56]. (b) 32 × 32 silicon E-O MZI-based Beneš switch. Reprinted from [57].

For fast E-O phase shifting, carrier-injection based PIN junctions are more widely applied in optical switching circuits than their carrier-depletion-type, PN counterparts. PIN junctions show higher efficiencies ($\partial\lambda / \partial V$) in terms of nm/V, effectively leading to a much smaller footprint and/or a lower operating voltage. The E-O MZI switch cell, however, requires additional design considerations as the induced electro-absorption loss imbalances power intensity in the two parallel arms, which deteriorates both crosstalk and insertion loss. 16 × 16 port count MZI-based silicon E-O switches were reported by two separate groups in 2016 [59,60]. Both demonstrations applied the Beneš architecture, which requires the least number of switch cells to offer non-blocking connectivity, thus relaxing the challenge for both design and fabrication. For a single MZI switch cell design, Lu *et al.* integrated a pair of TiN micro-heaters for both arms and a PIN diode for one arm to co-optimize device performance [59]. The use of a single E-O phase shifter, however, resulted in an evident discrepancy between the cross and bar states in loss and crosstalk due to the free carrier absorption. Qiao *et al.* implemented only PIN junctions for both arms but in an opposite set for push-pull switching operation [60]. The push-pull control scheme could somewhat mitigate the imbalance of power intensity in the two arms and more importantly, reduce the length of phase shifters (200 µm) due to the smaller phase change required. Power taps were introduced for both switch devices to help correct potential fabrication errors. Very recently, the connectivity of silicon E-O MZI-based switch has been scaled to 32 × 32 as a new record by Qiao *et al.* (Fig. 4b) [57]. This device also employed the Beneš topology implementing 144 MZI elements. An optical phase bias of π/2 was intentionally introduced to one phase-shifter arm of each MZI unit, in order to equalize the phase change of the two arms during push-pull operation. Therefore, the operation for the bar and cross state could be balanced and thus optimized. On-chip insertion losses and the crosstalk ratios were measured to be 12.9 to 16.5 dB and −17.9 to −24.8 dB, respectively when all the switch elements were configured to the cross state, while for the all-bar configuration, the numbers were in the range of 14.4 to 18.5 dB and −15.1 to −19.0 dB, respectively. It should be noted that the operation is not always optimal because of the fabrication errors. Further optimization can be achieved by adding an extra pair of T-O heaters to provide an additional degree of control in compensating fabrication variations [61]. We have demonstrated a fully-automated implementation that co-optimizes thermo-optic and electro-optic phase elements, correcting phase-error and power-imbalance simultaneously for optimized performance [61].

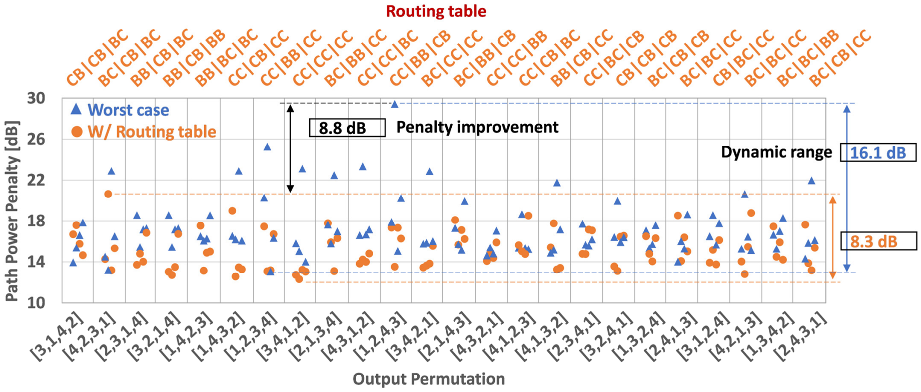

Fig. 5. Detailed port-to-port power penalty of both selected and worst-case path mapping for all 24 permutations. Note that 11 dB of power penalty contributed from in/out grating couplers. Reprinted from [68]

These E-O switch circuits indeed showed the feasibility of a high-level integration in silicon photonics; however, their current performance is far from practical for real applications. The intrinsic limitations on the performance and scalability are loss and crosstalk. Lower crosstalk can be obtained by topological modification to cancel the first-order crosstalk [56,58], but trading off a greater number of switching elements results in higher insertion loss. Studies were carried out to improve the performance of single MZI switch cells, and the introduction of optical phase bias and push-pull control is a notable example [57]. The application of broadband coupler [62] and nested MZI structure [63] were proposed for better crosstalk reduction but at the sacrifice of wavelength dependence. Later on, a balanced nested MZI switch structure was reported [64] offering broadband tri-state operation. The extra blocking state guaranteed an excellent crosstalk suppression. Insertion loss, however, is more challenging to manage for large-scale switch fabrics. SOAs are a natural solution to provide on-chip gain. A gain-integrated silicon photonic carrier prototype for silicon optical switches with a flip-chip bonded SOA array as a lumped amplification stage was developed [65]. The recent report of a lossless SOA-integrated 4 × 4 PILOSS silicon switch, leveraging the flip-chip bonding technique, was a notable demonstration [66].

In addition to the device-level optimization, we have further proposed a fabric-wide advanced routing method at the control domain, providing routing strategies that are aware of physical-layer performance and thus selecting the optimal solution. A comprehensive analysis on the Beneš topology was presented in [67]. The element of repetition between the global switching states (2^M, where M is total number of 2 × 2 switch elements) and the switch permutations (N!, where N = 2^m is the switch port count) is first quantitated with upper bound of $C_m(upper) = 2^{(m-1)2^{m-1}}$, and lower bound of $C_m(lower) = 2^{(2^{m-1}-1)}$, where m≥1. Path power penalties of any N paths that form a permutation in the switch fabric was simulated by aggregating insertion loss and crosstalk impairment. The results were subsequently sorted by their corresponding permutation and the deviation, i.e. root mean square error (RMSE), of power penalties for the categorized switching states was calculated. The one with the least value was stored as the routing strategy that equalized the path-dependent power penalty, generating a full look-up table for all permutations.

We also experimentally demonstrated the first full smart routing table in [68]. Each arm of the MZI module was equipped with a thermal tuner for device calibration and a PIN phase shifter for high-speed switching. By configuring each MZI element, the 4 × 4-port Beneš switch can be configured in 64 (2^6) global states that map to 24 (4!) permutations. By aggregating leakage crosstalk and translating into power penalty [69], the total path power penalty can be obtained for all 256 light paths corresponding to 64 switching states × 4 inputs. A complete look-up table can be generated, as shown in Fig. 5, as the solution that best equalizes path-dependent power penalties and avoids the worst-case path.

In addition, we have developed a highly accurate calibration and characterization process for optical switch fabrics without the need for built-in power monitors [70]. This technique can substantially reduce the cost and complexity for device integration and packaging, which can be leveraged to facilitate the fast generation of look-up tables benefitting from the fully automated process.

4.2 MRR-based optical switches

The first demonstration of µm-scale silicon MRR by Xu *et al.* has stimulated the research of MRR-based photonic integrated circuits [71]. Early research work on MRR-based switching circuits was led jointly by the teams from Columbia and Cornell [72–74]. A representative of the hitless router is shown in Fig. 6a [72]. The wavelength-selective nature of MRR unit does require wavelength alignment across the switching circuit, adding extra overhead. Various schemes for fast and efficient wavelength locking have been demonstrated [77–79]. Routing of WDM signals has been demonstrated leveraging the comb-switching technique [73]. Higher-order MRR elements can enable broadened passband with enhanced extinction ratio and thus relax the wavelength registration requirement [80], but at the cost of higher insertion loss and fabrication complexity. Fabricated devices so far tended to apply the crossbar architecture, which suits the add-drop feature of MRR cells. The largest port-count of this type of switch matrix reported to date is 8 × 7 based on thermally tuned fifth-order silicon MRRs (Fig. 6b) [75]. The single MRR switch element was designed to give a 100 GHz passband and a free spectral range (FSR) of 350 GHz. The on-state and off-state transfer functions revealed switch extinction ratios of better than −20 dB. The MRR switch element featured an averaged through loss of 0.9 dB and drop loss of 2.0 dB. The path-dependent loss was ranging from 14.5 to 22 dB due to different numbers of rings passed through. The performance and scalability of such a switch fabric is again limited by the insertion loss and extinction ratio.

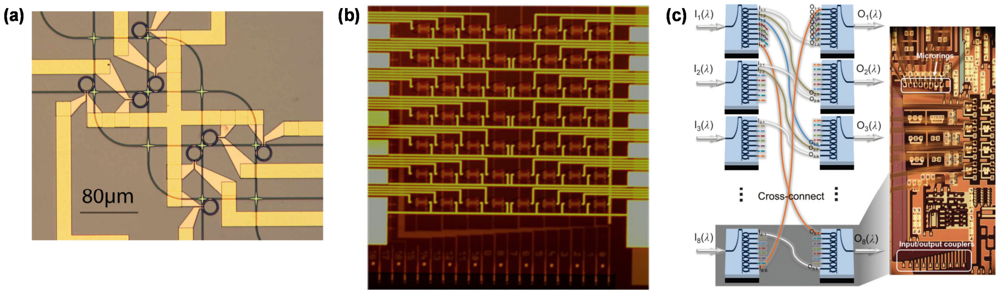

Fig. 6. (a) 4 × 4 silicon MRR-based hitless router. Reprinted from [72]. (b) 8 × 7 silicon T-O fifth-order-MRR switch. Reprinted from [75]. (c) Modular silicon switch-and-select MRR-based switch. Reprinted from [76].

Recently, we proposed a modular switch-and-select topology by assembling 1 × N/N × 1 ring-based spatial (de-)multiplexers with low-loss fibers or 2D optical interposer (Fig. 6c) [76]. Each (de-)multiplexer comprises N ring resonators coupled to a bus waveguide to add or drop optical signals. This design only allows second-order crosstalk and maintains the number of drop-rings per path at two, while further scale-up only adds bypassing rings through the bus. A proof-of-principle demonstration of an 8 × 8 switch-and-select MRR switch was performed with excellent results [76]. However, it scales with regard to the number of MRRs as $2N^2$ and for monolithic integration, managing waveguide crossings at the central shuffle network becomes more and more difficult at high numbers. Further study was performed in combining the scalable three-stage Clos network with populated switch-and-select stages [81]. This design offers a suitable balance that keeps the number of stages to the modest value of three while largely reduces the required number of switching elements, potentially yielding nanosecond-reconfigurable large-scale switch fabrics. The topological evaluation (Figs. 7(a) and 7(b)) by comparing the Clos-of-switch-and-select topology with other commonly applied architectures reveals that it becomes striking for large-scale networks in balancing the total number of rings and the level of cascading.

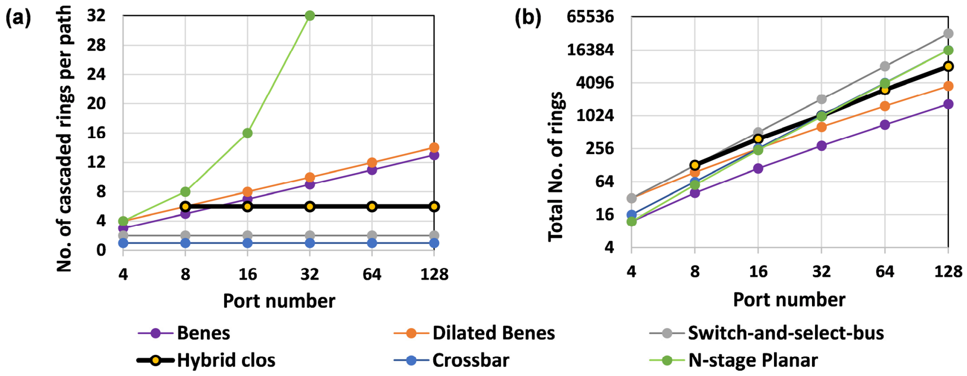

Fig. 7. (a) Number of worst-case drop rings per path and (b) total number of rings as a function of switch port number among different architectures.

The wavelength-selective nature of MRRs determines that they can be used for WSS applications with elastic WDM switching capability. The concept has been proposed by placing a number of silicon MRRs, with different resonating wavelengths, as wavelength-selective add-drop nodes in the crossbar topology [10].

4.3 Silicon integrated MEMS switch

A new type of monolithically integrated MEMS-based optical switch in the platform of silicon photonics has been demonstrated [82,83]. The switching cell is based on vertical adiabatic couplers actuated by MEMS elements. Switching elements, i.e. MZI or MRR, in conventional integrated switch fabrics are responsible for both propagating and redirecting the optical signal. This geometry induces loss and crosstalk at each switching stage and the accumulated signal impairment in turn limits the port count scalability. The silicon MEMS switch applied a dual-layer structure that decouples the functionality of signal redirecting from propagating at a switch node, and thus allows independent optimization. It is built based on the crossbar topology by a matrix of ultra-low loss waveguide crossings to form the lower-layer passive shuffle network for signal propagation. Signal redirection is performed by another layer of waveguide that is attached to a MEMS-actuated cantilever as movable adiabatic couplers. The coupling is electrically controlled by adjusting the vertical offset between the two layer waveguides. The application of cross-bar architecture ensures that the optical signal only passes the adiabatic coupler once; therefore, the cumulative loss and crosstalk can be significantly reduced and large-radix switch fabrics can be achieved.

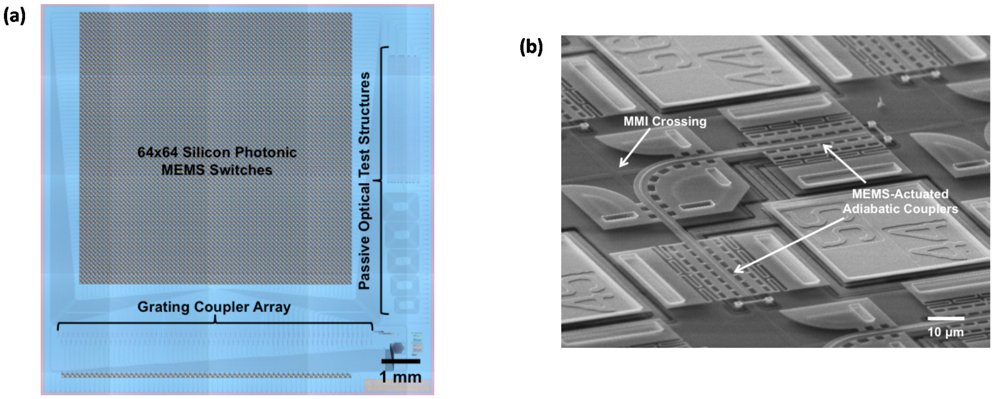
(a)
64x64 Silicon Photonic MEMS Switches
Passive Optical Test Structures
Grating Coupler Array
1 mm
(b)
MMI Crossing
MEMS-Actuated Adiabatic Couplers
10 μm

Fig. 8. (a) 64 × 64 silicon integrated MEMS switch. (b) MEMS switch cell. Reprinted from [83]

The MEMS-based silicon switch matrix has scaled to 64 × 64 port count (Fig. 8) [83], with 4096 switching cells in a compact footprint of 7 × 7 mm^2. This monolithically integrated switch fabric was fabricated on two stacked silicon layers: the first bus waveguide layer was patterned on the SOI wafer with a 220 nm thick silicon layer, and the second layer for adiabatic couplers and MEMS actuators was on the deposited polysilicon layer with a 300 nm thickness. MMI coupler was used to focus light at the center of the intersections in order to make the excess loss low, at 0.017 dB/crossing. Considering the waveguide propagation loss at 1.52 dB/cm, the through loss of a switch cell was estimated to be merely 0.028 dB with a footprint of 110 × 110 μm^2. The MEMS-actuated vertical coupler was designed to operate optimally at a gap spacing of 125 nm for the ON state and >800 nm for the OFF state. The coupling loss was measured to be 1.55 dB and the maximum insertion loss for the longest path of the 64 × 64 switch obtained was 5.1 dB. The ON/OFF extinction ratio was reported >65 dB for a single switch node. The MEMS actuation features an advantage of digital operation that eliminates the bias-dependent calibration. However, the turn-on voltage at 51 V is prohibitive for developing low-cost switching systems due to the complexity of drivers, though the electrostatic operation can be highly efficient. Sub-microsecond (0.91 μs) switching times were demonstrated.

5. Optical switching metrics for datacenters

In this section we discuss the applications of optical switching for computing and networking. As hinted in the previous section, performance and quality of switches are generally assessed in terms of scalability (number of ports), switching time, power penalty (loss and crosstalk), cost, and power consumption. A comparison among different optical switching technologies in the key figure of merits is summarized in Table 1. We in particular review the performance metrics that are required for optical switches to truly emerge in large scale computing systems such as datacenters.

Table 1. Comparison table of different optical switching technologies

	Scalability/ Record	Switching Time	Insertion Loss	Crosstalk	Power Consumption	Polarization
Free-space 3D MEMS	Very large/ >1000 × 1000	10-20 ms	Low	Low	Medium (high driving voltage)	Independent
LCOS	Medium/ 1 × 20	100 ms	Medium	Medium	Low	TE
III-V SOA	Medium/ 16 × 16, 8 × 8 × 8λ	~ns	Low even lossless	Low	High[a]	TE (can be independent)
III-V MZI	Small/ 8 × 8	~ns	High	High	Medium (PDE) Low (QCSE)	Independent
SiP T-O	Medium/ 32 × 32	10s of μs	Medium to large	Medium	High	TE (can be independent)
SiP E-O	Medium/ 32 × 32	~ns	Large	High	Medium (PDE)	TE (can be independent)
SiP MEMS	Large/ 64 × 64	Sub-μs	Medium	Low	Medium (high driving voltage)	TE (can be independent)

[a]Consumed power for both actuation and amplification

However, this list of metrics must be revised depending on the exact optical switching application. In this context, the importance of the mentioned switching metrics can be described as follows.

5.1 Optical switching time

The necessity for nanosecond or sub-nanosecond switching time is a frequent belief among the optics community. Yet, multiple system level facts can be brought to show that reducing the switching time below the 100 ns mark does not really bring any benefits. Considering today's bandwidth requirements of datacenter servers and supercomputer nodes, it is relatively clear that 100 Gb/s link bandwidth should be considered [15]. At the data lane level (i.e. per wavelength), transmission rates of at least 25 Gb/s are expected. At such speed, though, the links are extremely sensitive to phase variations (the duration of a bit being only 40 ps). Prior to sending any data on a 100 Gb/s data lane, a link training session must thus occur to understand the phase conditions resulting from the exact environment (fiber length, connectors, etc.) and adapt to it. Once this training session at the PHY level is over, the link must then be "locked" at the MAC layer and its clock aligned with one of the receivers. In other words, bringing a 100 Gb/s link to stable operational mode is a 10 to 100 ns process at least. Provided that reconfiguring an optical switch may relatively drastically change the propagation condition from a phase perspective, and also breaks prior links to provide new ones, after each modification of the configuration of an optical switch, a retraining of all the links traversing the switch is required. This retraining time takes, as already mentioned, certainly more than 10 ns and more likely more than 100 ns. Therefore, optical switching times less than 10 ns will be widely dominated by the training time, bringing no real benefits.

The need for link training is not the only fact that makes very short switching times not really desirable. In the context of using optical switches in addition to regular packet routers, changing the configuration of optical switches results in a modification of the network topology. This in turn requires a mandatory modification of the routing tables. As shown by Shen et al. [84], updating routing tables in OpenFlow type routers takes milliseconds. Even if equipment optimized for fast routing table update is developed, the fact that routing tables must be ideally all updated at the same time across the interconnects further complicates the operation and leads us to posit that topology reconfiguration time at the packet switch level takes several microseconds. We thus conclude that switching times below the microsecond mark are good enough for datacenter or supercomputer applications and that reducing switching times far below this mark will not yield big performance gains.

Switching times higher than a microsecond will not be dominated by link training and topology reconfiguration effects. However, such “high” switching times might not be problematic provided that reconfiguration of the optical switch occurs seldom enough. In [14], we have shown that in HPC, traffic flows tend to have very contrasted locality even over large periods of time. This means that even if topology reconfiguration operation occurs at the task granularity (i.e. every minute, hour, or day), substantial performance gains can be obtained. If an interconnect is reconfigured not more frequently than every hour, the negative impact of even second scale switching times will remain almost unnoticeable. The value of a switch as function of its switching time, all other metrics considered equal, can be thus considered dropping slowly in the 1 microsecond – 1 second range. Beyond 1 second, the value is expected to drop with a steeper slope, as illustrated in Fig. 9a and simplified in Fig. 9b.

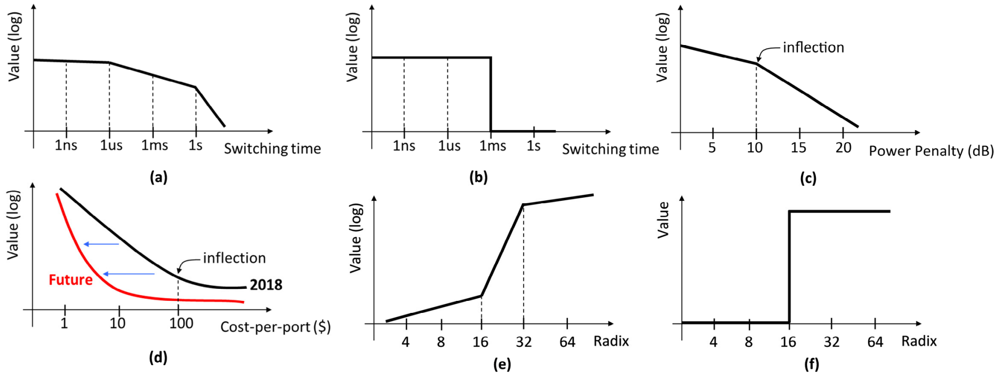

Fig. 9. (a) Relationship between switching time and value. (b) Step function estimation of value vs. switching time. (c) Relationship between switch power penalty (worst case) and value. (d) Relationship between switch cost-per-port and value. (e) Relationship between switch radix and value. (f) Proposed approximation for the value of the switch as a function of radix. In cases (a)-(f), all other parameters are considered unchanged.

5.2 Power penalty

In terms of power penalty, optical switches must fit within the optical budget of the transceivers. This is necessary both in terms of cost and power consumption. Hence, if one or more amplifiers must be added as discrete components to every link in a datacenter, the cost per link increases sensibly, possibly negating the benefits of optical switching. In terms of power consumption, important switch power penalties have an impact on the launch laser power. For instance, considering a receiver sensitivity of −16 dBm and a link budget of 20 dB, a typical launch power per wavelength is 4 dBm, i.e. 2.5 mW, which, for 25 Gb/s signaling per wavelength and 10% laser wall plug efficiency, corresponds to 1 pJ/bit. If a switch with a power penalty of 10 dB is inserted, not only must the laser provide an increased power of 14 dBm per wavelength, the energy per bit of the laser, in the absence of amplification, is raised by a factor of 10 to 10 pJ/bit. This is a lot, especially considering that electrical packet switches are now capable of receiving, switching and sending a bit while burning 20-30 pJ/bit. The presence of optical amplifiers can mitigate the increase of laser power consumption, yet the intrinsic power consumption of amplifiers must also be taken into account. Unless amplifiers with outstanding wall-plug efficiency can be developed, the impact of switch power penalty on the end-to-end link power consumption will remain important and negate energy efficiency benefits of optical switching, and might even threaten massive utilization of switches due to energy consumption concerns.

In summary, we expect the “value” of an optical switch, all other metrics assumed constant, to evolve with power penalty as described in Fig. 9c. For power penalties ranging from close to 0 dB to 10 dB, the value is generally “high”, and progressively increases as the power penalty decreases to 0 dB to reflect the energy efficiency.

5.3 Cost

The presence of optical switching in a network is essentially justified by the need to better use the limited resources available. By changing the topology using optical switches, pairs of servers or racks exchanging a lot of traffic can be awarded more bandwidth on demand. However, it is also possible to mitigate network congestion cases resulting from intense communications between servers or rack pairs by overprovisioning the network, i.e. by adding more bandwidth “just in case”. Let’s assume a system where a limited subset of the links are congested because the traffic that tries to transit over them is twice more intense than their bandwidth. Such a congestion case can be solved by reorganizing the bandwidth using optical switches, i.e. by requisitioning the links with no congestion and awarding their bandwidth to complement the one of the congested links [14]. However, such a congestion case can also be solved by simply doubling the number of links and the equipment present in the network. Of course, doubling the resources makes the resources’ cost to grow by more than a factor of two due to the overhead caused by extra complexity. However, the increase in the cost is not expected to be more than a factor of ~3.

This reasoning leads us to posit that for optical switching to be competitive, the cost associated with making a link configurable must be considerably smaller than the cost of adding an additional link. For simplicity, let’s assume that “considerably” means one order of magnitude. Based on this assumption, we can derive a desirable cost per port figure. 100 Gb/s optical links currently come at a price on the order of $1,000 (one active optical cable or one pair of transceivers + fiber, and two switch ports) [85]. To be cost competitive, optical switches should thus currently be sold at a cost per port on the order of $100/port. This corresponds to a price of $32,000 for a 320 port optical switch as the one proposed by Calient [27]. However, as the link speeds progressively evolve from 100 Gb/s to higher values, the proportion of optical links in a large-scale interconnect will increase (because the fewer links will be short enough to support 400 Gb/s in copper). This will put some pressure on the cost of optical links, obliging them to be sold at a price closer to the one of copper links. One can thus expect the cost per link to decrease to a value closer to $100 [15]. To keep one order of magnitude advantage, optical switches should thus be proposed at a cost per port of ~$10.

Figure 9(d), similar to Figs. 9(a) and 9(c), relates how the cost-per-port is impacting the switch value. Value can be considered very small as long as the cost-per-port is more than $100 (in 2018). As this cost is reduced, the benefits of optical switching compared to overprovisioning grows and so does the switch value. We expect the curve to push to the left over time, as illustrated by the thin arrows.

5.4 Electrical power consumption

Optical switches indirectly affect power consumption through optical power penalty, but also have their own intrinsic power consumption. We note that assuming an electricity price of 0.1 $/kWh, and a system (datacenter, supercomputer) lifetime of 5 years, 1 W of power consumption costs 365 × 5 × 24 × 0.1 / 1000 = 4.38 $, a value that can be rounded up to $10 to take into account power supply overheads. If a port consumes 1mW, a negligible 1 cent must be added to the cost-per-port as defined above, but if a port consumes 1 W, the associated cost must be raised by $10. Therefore, the electrical power consumption should be taken into account as a contributor to the cost.

5.5 Radix/Scalability

At equal cost per port, switches offering the highest number of ports will obviously be privileged over low port count switches, as the former will enable more flexibility in the way bandwidth can be reorganized. However, we have shown in [15] that the benefits of being able to reorganize a network can be reaped even with modest radix switches, the break-even point lying probably around 8 to 16 ports. This value might end up being higher for datacenters but probably not higher than 32 ports. In other words, it is possible to design an interconnect with the required reconfigurability potential with 32 ports switches. Elaborating such designs will require more attention from the designer, and a deeper knowledge of the traffic patterns that are expected. Yet, figuring out methods to correctly place modest radix switches (as the prototype one we proposed in [15]) is a one-time effort.

This reasoning leads us to conclude that for applications in large scale datacenter interconnects, optical switches must propose a minimal number of 16 to 32 ports. Beyond this minimal requirement, the increase in value becomes marginal, as illustrated in Fig. 9e.

5.6 Combining metrics

Out of the metrics introduced so far, the relationship to value can be simplified in two cases. First, in terms of radix, we think that we can approximate the relationship sketched in Fig. 9e by a step function as shown in Fig. 9f. The value of a switch is deemed null for radixes less than 16 ports. Beyond this radix, the value is considered constant. This approximation might look outrageous, but essentially by making it, we consider that increasing radix provides second order benefits compared to the other metrics. Second, we think that the switching time can similarly be presented by a step function: any switch whose switching time is less than or equal to the millisecond mark is considered valuable. This generally qualifies most optical switching technologies, thus excluding switching time as a relevant metric for assessing the switch value.

Having made these approximations, the space explored by the metrics presented so far is reduced to a cost-per-port/power penalty 2D plane, as represented in Fig. 10a. We can then approximately spot the four representative technologies on this plane, i.e. free-space MEMS switch, III-V SOA switch, SiP MEMS switch and SiP interferometric switch. The data quoted for optical power penalties were acquired from the state-of-the-art demonstrations for each switching technology [12,25,83,86]. For cost-per-port, we set the free-space MEMS switch and SiP interferometric switch as two reference points. The free-space MEMS switch is currently priced at $100-200 per port [27] and considering the rigorous calibration and installation of discrete components, e.g. 2D fiber arrays, micro-lens arrays and 3D beam steering mirrors, and the relatively narrow production volume, it is unlikely to have a fully automated machine to eliminate the labor cost and thus tremendously drives the price down in the near term. From there stems the ~$100/port estimate. As for the silicon photonics, the packaging cost currently dominates such as the case for optical transceivers, but considering the current research effort to realize low-cost fiber attachment (for both transceivers and switches) [87], we posit that packaging cost will be drastically reduced in the next years, since the packaging costs must be rationalized to reach the frequently noted cost target of <1$/Gb/s [88]. We thus expect the packaging of a chip-integrated switch to become of the same complexity as the one of a modern CPU with dense IOs. Provided that the price of Intel CPUs using old fabrication nodes, such as Celeron, has dropped to a few tens of dollars, we conclude that the cost-per-port for SiP interferometric switches can be potentially at ~1$. For silicon MEMS switches, the cost premium compared to SiP interferometric device is assumed principally due to the high driving voltages that require specialty circuits possibly not feasible in CMOS. The additional fabrication process, for instance depositing polysilicon layer for adiabatic couplers and MEMS actuators, is another consideration. The difference in wafer cost and available wafer size for InP and SOI is a notable factor when comparing the cost-per-port between SOA-based switches and SiP interferometric ones. In addition, the multiple epitaxial growth processes for InP-based devices could be a threat to the yield. These factors can certainly be overset by the market volume in the future. It can be seen in Fig. 10a that free-space MEMS switch shows very good power penalty but high cost. SiP interferometric switches (MRR or MZI based) have the potential for ultra-low cost per port, but suffer from high power penalties. III-V SOA integrated switches are more expensive than silicon devices, but the lossless capability largely brings down the power penalty and the limitation depends on the noise and distortion. Silicon integrated MEMS switches show good balance between the cost and power penalty. It should be noted that continuing efforts are being poured into the study of photonic switches to push the technologies towards the lower left corner, and with future advancement in device fabrication, coupling, packaging and assembling, the Pareto front will change.

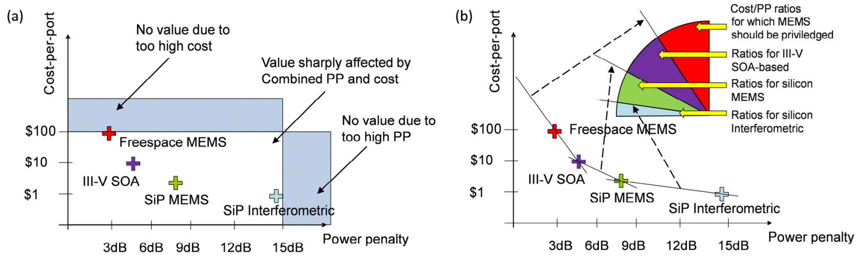

Fig. 10. (a) 2D plane Pareto analysis of switch quality. (b) Pareto analysis of four switching technologies. The space of each element in the pie indicates the ratios that a switching technology should be privileged as a combining factor of the cost-per-port and power penalty.

We see that four technologies, i.e. free-space MEMS, silicon integrated MEMS, III-V SOA and SiP interferometric switches, diversely populate the Pareto front (Fig. 10b) and that no one fully dominates the others. Free-space MEMS should be favored if power penalty is of paramount importance compared to cost. This can be the case if the laser efficiency is not very good and if any link power penalty would result in heavy electricity cost overhead (i.e. because of power penalty, the laser consumes much more, which eventually makes energy to cost more than the ports). In return, SiP interferometric switches should be selected if cost is of most importance. Applying the same reasoning, that could be the case when laser efficiency is very good and power penalty marginally affects the total cost of ownership through electricity consumption. For cases in between, the III-V SOA and silicon integrated MEMS should be favored. In light of these analyses, we further introduce in the next section a switching technology that combines SiP interferometric switching elements and III-V SOAs.

6. Hybrid concept: III-V/Si switches

The advancement on wafer-bonding technique stimulates a new class of integrated devices [89], in which III-V layers are bonded on silicon wafers. The bonding can either be direct [89] or adhesive with polymers [90]. This effectively combines the strong points of both platforms and potentially offers batch fabrication. The enabled hybrid concept could potentially offer a new opportunity to provide compact, energy efficient and low-cost switch fabrics satisfying the metrics stated in the last section.

The hybrid concept for optical switching via wafer-bonding was previously explored by combining the III-V material with SOI platform to enhance the electro-optic effect for phase manipulation [91]. To our best knowledge, adding III-V gain block to silicon switch circuits so far were mainly carried out by the flip-chip bonding approach [65,66,92]. The flip-chip bonding technique, however, requires precise control on the coupling between two materials and low reflection at both facets, and the assembling may counteract the low-cost benefit of silicon fabrication. Standalone hybrid SOA element by direct wafer-bonding was first demonstrated more than a decade ago, in which the optical mode is laterally confined by the silicon waveguide and couples evanescently to the III-V layer [93]. This configuration features less complexity in the inter-layer coupling, but the optical mode only interacts with the III-V slab via its evanescent tail and thus has a restricted modal gain. Thickness of the bonding layer is also critical to the amplification. Tapered structure is required for the III-V mesa to increase coupling efficiency and minimize reflection. By varying the silicon waveguide width to adjust the confinement in the active layer, a trade-off can be decided between the optical gain and saturation power [94]. The restricted modal gain, however, calls for long SOAs. The later work proposed to have the optical mode reside in the III-V waveguide to experience the maximal modal gain with adiabatic mode transformers between the two materials for hybrid lasers [95]. This design can in principle be as efficient as a monolithic InP laser diode and can be directly adopted for hybrid amplifier design in optical switching circuits. A high-efficiency but low-reflection power transfer between the III-V mesa and silicon waveguide requires careful examination on the adiabatic coupler design [96]. The complexity in the adiabatic mode transformers could be a threat to the device yield, and this inevitably leads to a large footprint.

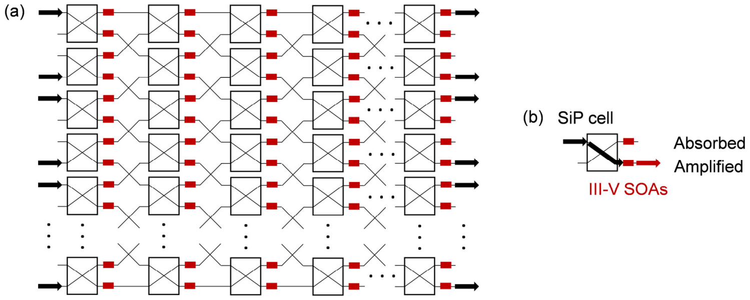

Fig. 11. (a) Schematic of a hybrid III/V-Si switch in a PILOSS topology. (b) A hybrid switching cell carrying one input signal at once and the attached SOAs switched in tandem.

The hybrid design targets lossless operation that is significantly advantageous within datacenters because this not only avoids additional optical amplification such as EFDAs, but also allows the optical transmitters to operate at moderate output power and remove the need for extensive electrical amplification at the receiver side. We would like to introduce SOA elements in a distributed way such that the absorption can be additionally used for crosstalk suppression, in a similar concept as the III-V integrated switches [49]. The difference is the interferometric switching elements are implemented on silicon substrate and either MRR or MZI cells can be adopted. For switch topologies that cancel the first-order crosstalk, each switching cell only carries one signal at once and SOAs are switched in tandem, amplifying in on-state while absorbing in off-state. The dilated Beneš [49] and the custom dilated non-blocking switch [58] are good candidates. PILOSS is another candidate and the uniform distribution of switching cells together with waveguide shuffles is a big bonus for SOA-involved multi-stage switch design (Fig. 11), as discussed in section *3.1*. However, it should be noted that the PILOSS topology does not completely block first-order crosstalk [97]. In addition, the attached SOAs can operate as power detectors in the reverse bias mode for a self-calibration purpose [98]. Given that power-efficient optical switching has been demonstrated for both electro-optic [57] and thermo-optic [86] devices on SOI platform, the energy efficiency of III-V/Si hybrid switches will largely depend on the coupling efficiency between the two materials. The demonstrated high wall-plug efficiency of the hybrid distributed feedback laser paved the way for power-efficient hybrid switch devices [96].

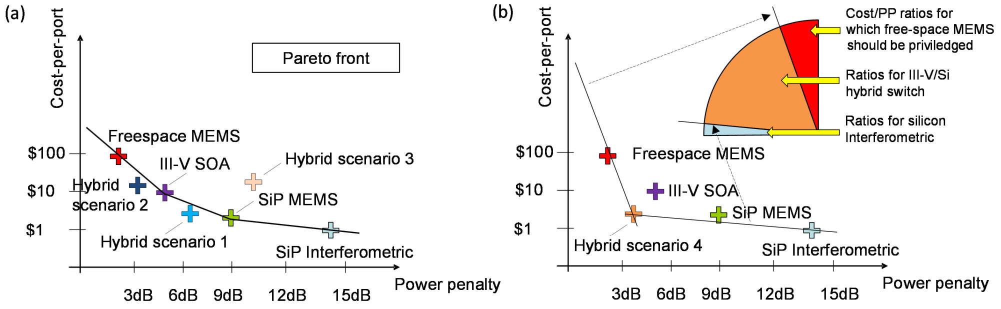

Fig. 12. Potential impact of hybrid integrated switch fabrics for (a) scenario 1, 2, and (b) 3 by inserting the III-V/Si hybrid switch in the Pareto front.

We now provide a discussion on the potential impact of the hybrid integrated switch fabric and insert it in the metric space. In the first scenario (marked in Fig. 12a), the power penalty is largely reduced comparing to SiP interferometric ones due to the additional gain and crosstalk suppression, but the technology is not matured to make it close to the free-space MEMS counterparts. On the other hand, the cost impact of the presence of SOA can be minimized, as it is envisioned that by wafer bonding and hydrogen-induced layer transfer, a single InP substrate could be reused and thus reduce the cost [99]. This permits the hybrid integrated switches to be useful for cases where cost is of high importance. In the second scenario (marked in Fig. 12a), as the hybrid integration technology is maturing, high performance in regards of the power penalty can be obtained. Therefore, increased fabrication costs can be tolerated. In both cases, the hybrid device will populate in the Pareto front. The third scenario (marked in Fig. 12a) represents the worst case that both the cost and performance of the hybrid integrated device fall behind comparing to other competing technologies. The ideal case, i.e. scenario 4, is presented in Fig. 12b, in which the hybrid integrated switch fabric is at a comparable price-level to SiP interferometric devices and exhibits power penalties close to that of free-space switch fabrics, becoming the most favorable choice.

7. Summary

The ever-growing interconnect demand of datacenters motivates the deployment of new technologies. Optical switching has received much attention to potentially address the challenges in regards to the bandwidth, cost and power consumption in datacenters. In this work, we reviewed the state-of-the-art MEMS, LCOS, SOA, MZI and MRR switching technologies from three optical switching platforms including free-space, III-V photonic integration and silicon photonic integration. Key figure of merits in terms of port count, switching time, power consumption, and optical power penalty are highlighted for datacenter applications. Furthermore, an evaluation methodology was introduced to assess the performance metrics required for optical switches and the Pareto front of the switch metric space was identified. It was shown that the free-space MEMS, silicon integrated MEMS, III-V SOA-based and SiP interferometric switches diversely populate the Pareto front, each playing a role for different application scenarios, and no one fully dominates the others. Finally, a hybrid III-V/Si optical switch was envisioned leveraging the III-V/Si wafer bonding technique to combine the strong points of both platforms. Its potential impact on realizing low cost and high performance switch fabrics for datacenters was investigated and discussed.

Funding

Air Force Research Laboratory (FA8650-15-2-5220); ARPA-E ENLITENED project (DE-AR00000843).