SAND2018-2328R

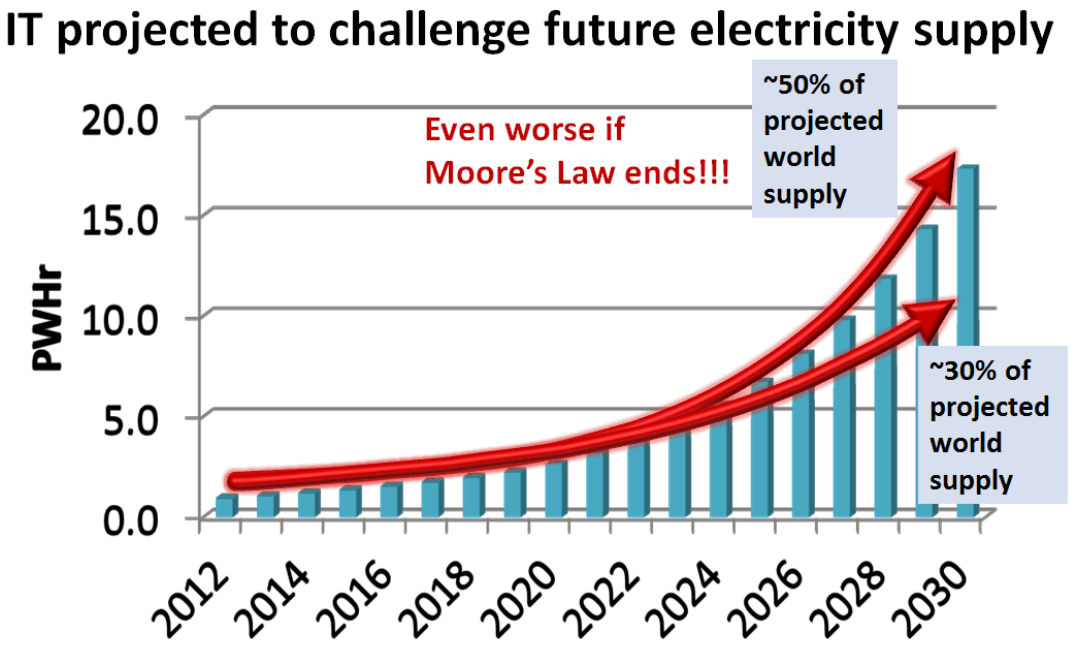
This report captures the initial conclusions of the DOE seven National Lab team collaborating on the "Solving the Information Technology Energy Challenge Beyond Moore's Law" initiative from the DOE Big Idea Summit III held in April of 2016. The seven Labs held a workshop in Albuquerque, NM in late July 2016 and gathered 40 researchers into 5 working groups: 4 groups spanning the levels of the co-design framework shown below, and a 5th working group focused on extending and advancing manufacturing approaches and coupling their constraints to all of the framework levels. These working groups have identified unique capabilities within the Labs to support the key challenges of this Beyond Moore's Law Computing (BMC) vision, as well as example first steps and potential roadmaps for technology development.

Executive Summary

A Looming Crisis: A multi-pronged crisis is threatening the $3.5 trillion-per-year global IT market that supports almost every aspect of our daily lives.[1] Over the next decade, our energy independence,[2] our global economic competitiveness,[3] and our national security[4] will be put at risk by the challenges facing the semiconductor and computing industry. Over the last five decades, Moore's law scaling turned computing into a pervasive consumer technology that steadily and reliably became increasingly more powerful and created the exponential growth of today's information economy. Moore's Law is a techno-economic model that has enabled the Information Technology (IT) industry to nearly double the performance and functionality of digital electronics roughly every two years within a fixed cost and area. Over most of these decades, Dennard scaling,[5] or Koomey's Law,[6] has provided this doubling of computing performance *without*

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

increasing energy consumption.[7] For 50 years, we have enjoyed the fruits of this "virtuous cycle" [8] in which investment enables a technology scaling increment leading to a performance increment that enables a cost-effective "killer app" (calculators, Walkman, desktop computers (PCs) laptops, iPods, iPhones, etc.) that generates enough revenue for the next cycle's larger investment. After dozens of iterations of this cycle, society has come to expect and rely upon the cost, energy efficiency, and storage capacity improvements provided by Moore's and Koomey's Laws.[9] However, we saw the end of Dennard scaling around 2006, and within a decade, the technological and economic underpinnings for the process Moore described will come to an end. At that point, it will be feasible to create lithographically produced devices with characteristic dimensions in the 3nm–5nm range. This range corresponds to about a dozen atoms across critical device features and will therefore be a practical limit for controlling charge in a classical sense. The classical technological driver that has underpinned Moore's law for the past 50 years is already failing and is anticipated to flatten by 2025.[10] The difficulty of pursuing these extraordinary limits of precision will drive the price of a leading edge fab in 2020 to over $15B,[11] causing extensive consolidation and mergers of companies in the semiconductor ecosystem.[12] Tapering of chip performance improvements in the coming decade will greatly accelerate the energy consumption of Information Technology (already the fastest growing consumer of energy worldwide) and/or significantly restrict U.S. computing growth, severely threatening the nation's ability to solve important problems in science, manufacturing, and national security.[4] A transition to a new scaling path will require effort on a decadal time scale, so it is critical to be laying the strategic foundation for change now.



www.alliancetrustinvestments.com/sri-hub/posts/Energy-efficient-data-centres
www.iea.org/publications/freepublications/publication/gigawatts2009.pdf

The dependence of our economic competitiveness on computing cannot be overstated. Almost half of the $3.5 trillion-per-year [13] IT market is spent in the manufacturing, finance, communication and government sectors, and high performance computing and mobile computing is now part of life and business at every scale:[14] social media, speech and image recognition, advanced design

and manufacturing, cyber-security, artificial intelligence, etc. Failure to keep up in computing means failure to participate and/or lead in the future economy. Computing innovation provides the agility needed to thrive in the information economy. Unfortunately, innovation is tightly coupled to manufacturing and the consolidation of the semiconductor microelectronics industry and movement of semiconductor manufacturing offshore is eroding our innovation leadership.[15] A concerted effort to position the U.S. as a leader in the post-Moore's Law future is needed.

Our national security is similarly dependent on computing. Economic and energy security are key parts of our overall national security, but "information dominance" has become critical to intelligence and cyber security activities, to design, build, and control military hardware, and to execute military actions in an effective manner.[16-17]

An Urgent Need for Action: The 50-year sustained growth of computing capability at relatively low consumer cost in terms of both dollars and energy has masked the urgency of our current situation. The energy timeline is urgent: a simple analysis of the growth of computing consumption versus performance scaling indicates that global electricity consumption by the information technology (IT) sector may need to grow from its current 3-4% fraction to over 30% in the next two decades to support the progress we've come to expect. This sector represents the fastest growing consumer of energy, and uncontrolled, this demand would have significant implications on the U.S. energy landscape. In one example of projected IT energy growth, Cisco reports that data center traffic (a useful metric for energy demand) is projected to grow at a compound annual rate (CAGR) of 25% from 2014-19.[18] With no improvement in computing efficiency (i.e., the end of Moore's Law), one would expect this growth to be directly reflected in increased energy demand going from 91 billion kilowatt-hours in 2013 to 252 billion kilowatt-hours in 2018 out of a total U.S. consumption of over 4000 billion kWhrs.[6, 18] Just meeting this increased demand alone would require 60 new 500-megawatt power plants, and the demand will likely be exacerbated by the end of conventional Moore's Law technology scaling in the next decade.

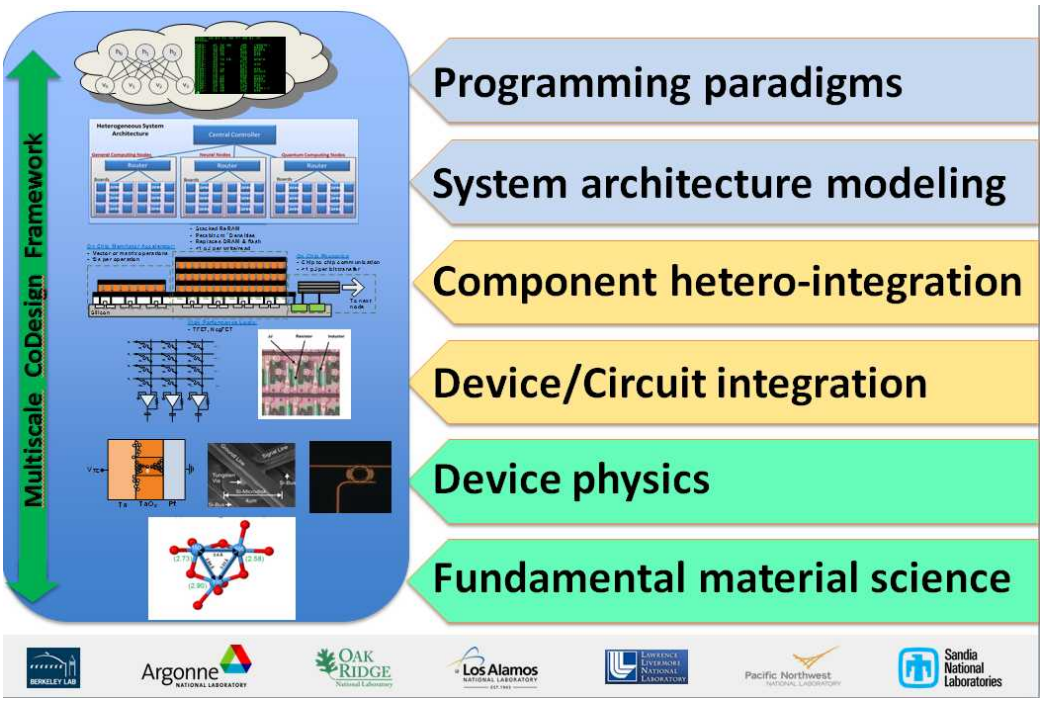
Economic and national security timelines are also urgent and disruptions in the pace of our progress in computing can have dramatic consequences: financial management, advanced manufacturing, online sales, global supply chain management are all demanding greater information processing fidelity at higher speeds, and the consequences of missteps are increasing (witness recent layoffs and increased consolidation of the entire ecosystem). A similar situation exists in our military and intelligence missions against terrorism, cyber-warfare, and modern ground and air warfare, where the pace of strategic and tactical decision-making is rapidly accelerating.

A Public-Private Partnership: The U.S. has a unique opportunity to create a new Public-Private partnership for basic/applied research to accelerate the development of energy efficient IT beyond the end of current CMOS roadmaps, as well as maintain an advanced manufacturing base in the economically critical semiconductor space.[2] This partnership will enable the government to coordinate current and future BMC investments more effectively, as well as to leverage significant industry investments. Pieces of this partnership already exist, in programs within multiple DOE and NSF program offices, within IARPA and NSA, and at NIST and DARPA. A large number of university projects exist in this area, supported by DOE and NSF funding, and at the Semiconductor Research Corporation (SRC) using DARPA, NSF, and Industry funding.

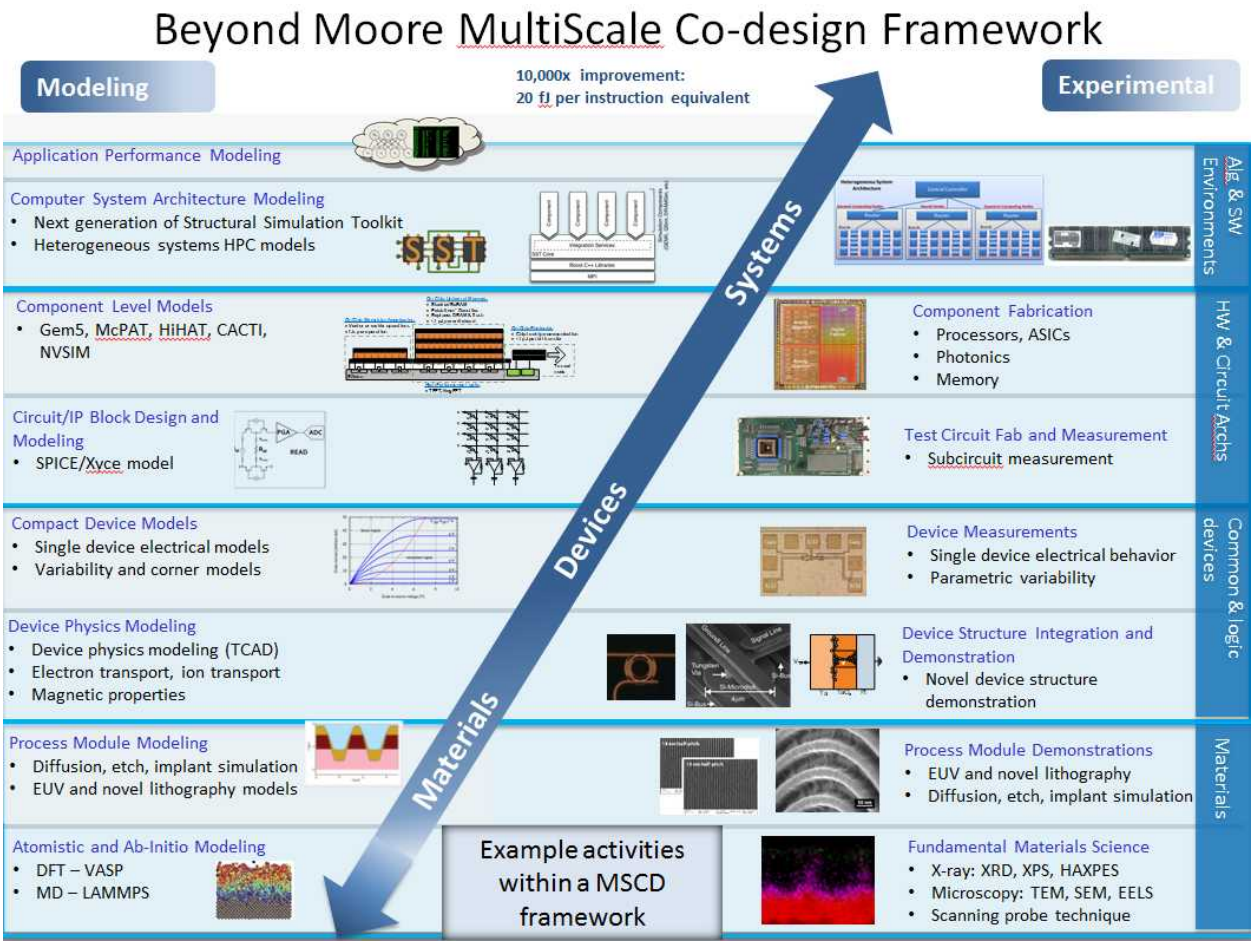
The Department of Energy (DOE) is well positioned to help address the developing threats, with program offices and national labs that serve the nation in all three of these problem areas (energy, economy, and national security), and a wealth of unique capabilities and depth of useful expertise. DOE Labs already support some of the ongoing work listed above, but largely in a diffuse, "individual relationship" manner. A broader initiative would provide easier access to the wealth of subsidized capability at the Labs, as well as break down some stovepipes and replicated efforts. The DOE also has a leadership role in the National Strategic Computing Initiative, is a primary driver for the "top of the computing pyramid" (HPC), and has successful experience with energy efficiency initiatives. The bulk of the impact of a successful initiative in this area would be achieved by increasing the efficiencies and manufacturing base for mobile, consumer, and "commodity" computing rather than "bleeding–edge" HPC, but advances from HPC investments within DOE will drive computing hardware and software progress forward, such that leading edge HPC capability becomes commodity more rapidly. Additionally, the science performed on, and enabled by, HPC at the labs benefits industrial breakthroughs and spin-off support technologies.

Fortunately, the U.S. semiconductor industry is still healthy, and well-positioned to continue to dominate the "more-Moore's Law" battle of the next decade. U.S. semiconductor companies account for about 50% of the global market share (2014) and still do over 50% of their manufacturing in the United States.[19] Semiconductors are one of America's top manufactured exports, behind only aircraft and automobiles.[20] While industry is currently focused on the intensely competitive challenges of "more-Moore's Law," a "wild zoo" of innovative R&D is growing in the Beyond Moore's Law computing (BMC) arena, across new materials, devices, manufacturing approaches, components, architectures, and algorithms.[21] While impressive and innovative R&D is underway, much of it is diffuse and uncoordinated—leading to an "organic evolution" of ideas and technologies, and slow progress towards a new path to scaling.

Solving the daunting technology, economic competitiveness, and energy challenges described above will require coordinated and coupled advances in materials and device technology going beyond CMOS, in scalable circuit integration, manufacturing, and packaging technologies, and in novel system architectures and programming models. ***Our team of eight (8) DOE national labs is proposing a Multiscale Co-Design Framework that will enable both top-down coupling of application and architectural requirements to circuits and devices, as well as bottom up coupling of materials and device physics constraints to algorithms and architectures.*** Portions of this sort of framework exist in today's industrial toolkit, but they are narrowly focused on the current Silicon technology stack, and it is very challenging to tightly couple activities across these different levels and technical disciplines. Additionally, the time and expense to develop strong coupling is poorly matched to the rapid pace of industry development cycles.



A Co-Design Approach: Our vision is for a Multiscale Co-Design Framework that will couple modelling at all of the key technology levels, validated with experimental data, parameter extraction, and physical demonstrations. This approach will exploit a large base of multiscale, multiphysics modelling and HPC at the Labs, as well as the large unique base of experimental, fabrication, and metrology capabilities maintained across the nation within these facilities. Indeed, small-scale versions of this approach are already underway at the Labs (and universities). It will also leverage the broad and deep expertise of Labs scientists, and the "shared culture" of the DOE Labs to overcome communication hurdles encountered in working across the technical disciplines within the layers of the framework.



Such a framework will provide a means to identify the most promising ideas developed by the research community through the lens of application impact and to avoid some of the "blind alleys" (material incompatibilities, un-manufacturable devices, circuit resiliency, etc.). It will also enhance communication and coordination across disciplines and organizations in industry, academia, and government. This co-design framework can eventually deploy powerful optimization and sensitivity/variance analysis codes to evaluate design tradeoffs across many levels and provide a means to develop "educated bets," quantify potential gains, and measure progress in this multi-disciplinary R&D environment. Only with co-design covering this broad space and consideration of manufacturing challenges can we expect to make progress in all areas cohesively to bring about real change to the IT energy outlook. In addition to containing the growth of IT-related energy demand, the output of this work will provide a path to sustaining exponential growth in computing capabilities to enable new scientific discoveries, and maintain U.S. competitiveness in all segments of the computing market (from IoT, to datacenters, to supercomputing), and ensure U.S. economic competitiveness and national security.

To meet the goals of broad societal impact, we must ensure transition of basic research to high volume manufacturing, and even more fundamentally shape basic research from the start with an eye to manufacturability. This will be achieved through the development of a multi-lab ecosystem serving as a facility that can evaluate and demonstrate the manufacturing and energy savings

feasibility of next generation technology options. Technologies will be rigorously evaluated for potential benefits on energy and implications on programming paradigms. The most promising technologies will be evaluated for high volume manufacturing feasibility followed by ramp-up demonstration and validation of energy impact. This phase will depend heavily on leveraging HPC capabilities to accelerate the development through modeling and "virtual cycles of learning."

A Multi-Agency Public-Private Partnership (PPP) Solution: Society will continue to rely heavily on the benefits of computing technology scaling. Evolving technology in the Moore's law vacuum will require an investment now to study candidate replacement materials, alternative device physics, circuit integration and manufacturing, architectures, etc. to foster continued technology scaling. Options abound, the race outcome is undecided, and the prize is invaluable. The winner not only will influence chip technology, but will define a new direction for the entire computing industry.

To avoid disruption or slowing of our information economy leadership, we must develop a National-level approach to accelerate our discovery and application timelines. The silicon FinFET is a relatively recent addition to Moore's Law, and while it uses much the same materials, physics, and manufacturing as prior device approaches, it took over 10 years for this advance in device physics to reach mainstream use. A more disruptive new technology might require even longer lead times and sustained R&D of one to two decades. To shorten this timeline, we seek a coordinated multi-agency effort that would:

1. Strengthen current and build new DOE-Lab partnerships with the semiconductor industry to continue Moore's Law in the near term and set the stage for a seamless transition to a "more than Moore's Law" landscape: more efficient Silicon CMOS electronics plus new accelerator technologies to enable computing past exascale.
2. Accelerate the maturation and transition of university R&D to industry.
3. Accelerate the development of a new energy/performance "scaling law" based on new BMC devices and architectures and the required new manufacturing infrastructure and ecosystem.

A first step towards this goal could be the creation of a network of 3-4 national computing technology centers focused on broad exemplar applications (science/physics code applications, data/search code applications, neuromorphic/brain inspired code applications). These centers will develop and execute the multiscale co-design framework, sharing capabilities, key elements, lessons learned, and best practices across the Lab, University, and Industry partners in all the centers. A multi-agency leadership team under the NSCI auspices will coordinate a Government + University + Industry "Manhattan Project" to cross-cut stovepipes within DOE programs, between Labs, and across Agencies. The funding structure for such a PPP might resemble:

1. $100-200M/year/Center
2. 50%:50% government:industry funding for years 1-3
3. 40%:60% government:industry funding for years 4-6
4. 20%:80% government:industry funding for years 8-10
5. 10%:90% government:industry funding for years 10+

REFERENCES

[1] "Gartner IT Forecast," http://www.gartner.com/technology/research/it-spending-forecast/.

[2] "Semiconductor Industry Association and the Semiconductor Research Corporation," **September 2016**, http://www.semiconductors.org/clientuploads/Resources/RITR%20WEB%20version%20FINAL.pdf.

[3] D. Rohrabacher, "Congress Member Dana Rohrabacher letter concerning foreign investment in U.S. semiconductor industry," Letter written June 22, 2015 to Treasury Secretary Jacob Lew, **June 22, 2015**, www.rohrabacher.house.gov.

[4] "American-Made National Security. (n.d.)," http://www.americanmanufacturing.org/campaign/entry/national-security

[5] J. M. Shalf, R. Leland, "Computing beyond Moore's Law," *IEEE Computer* **2015**, *48*, 14-23.

[6] J. B. Koomey, Stephen; Sanchez, Marla; Wong, Henry, "Implications of Historical Trends in the Electrical Efficiency of Computing," *IEEE Annals of the History of Computing* **29 March 2010**, *33*, 46-54.

[7] S. Lemon, T. Krazit, "With chips, Moore's Law is not the problem," **April 19, 2005**, http://www.infoworld.com/article/2669732/computer-hardware/with-chips--moore-s-law-is-not-the-problem.html.

[8] B. Krzanich, "Our Strategy and The Future of Intel," **April 26, 2016**, https://newsroom.intel.com/editorials/brian-krzanich-our-strategy-and-the-future-of-intel/

[9] M. M. Waldrop, "The chips are down for Moore's law," **February 9, 2016**, http://www.nature.com/news/the-chips-are-down-for-moore-s-law-1.19338

[10] "International Technology Roadmap for Semiconductors 2.0 2015 Edition, Executive Report," International Technology Roadmap for Semiconductors, **2015**, https://www.semiconductors.org/clientuploads/Research_Technology/ITRS/2015/0_2015%20ITRS%202.0%20Executive%20Report%20(1).pdf

[11] N. Mokhoff, "Semi industry fab costs limit industry growth," **October 3, 2012**, http://www.eetimes.com/document.asp?doc_id=1264577.

[12] D. Ford, "Preliminary 2015 Semiconductor Market Shares - IHS Technology," **December 8, 2015**, https://technology.ihs.com/553230/preliminary-2015-semiconductor-market-shares

[13] "Gartner Worldwide IT spending Forecast," http://www.gartner.com/technology/research/it-spending-forecast/.

[14] "Accelerance," http://www.accelerance.com/research/global-it-market-size-facts-and-figures#Summary

[15] D. Steinbock, " Overcoming the Erosion of American Innovation," **March 13, 2015**, http://intpolicydigest.org/2015/03/13/overcoming-the-erosion-of-american-innovation/.

[16] "Department of Defense Annual Report to Congress: Military and Security Developments Involving the People's Republic of China," Office of the Secretary of Defense, **2013**, https://www.defense.gov/Portals/1/Documents/pubs/2015_China_Military_Power_Report.pdf

[17] "U.S. Navy Information Dominance Roadmap 2013-2028," U.S. Navy, **2013**, http://www.defenseinnovationmarketplace.mil/resources/Information_Dominance_Roadmap_March_2013.pdf

[18] Cisco, "Cisco Global Cloud Index: Forecast and Methodology, 2014–2019 White Paper," **April 21, 2016**, http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.html?CAMPAIGN=GCI 2014

[19] F. Yinug, "Made in America: The Facts about Semiconductor Manufacturing," Semiconductor Industry Association, **August 2015**, http://www.semiconductors.org/clientuploads/Industry%20Statistics/FINAL%20Made%20in%20America%20white%20paper%20-%2008192015%20as%20posted.pdf

[20] "The U.S. Semiconductor Industry: 2015 Factbook," Semiconductor Industry Association, **2015**, http://www.semiconductors.org/clientuploads/Industry%20Statistics/2015%20Factbook/2015%20Factbook%20-%20Complete%20Updated%20-%2008062015.pdf

[21] B. Bailey, "Predictions For 2016: Semiconductors, Manufacturing And Design," **January 21, 2016**, http://semiengineering.com/predictions-for-2016-semiconductors-manufacturing-and-design/.