Testimony of Mark P. Mills
Executive Director, National Center for Energy Analytics
Distinguished Senior Fellow, Texas Public Policy Foundation
Before the
U.S. Senate Committee on Energy and Natural Resources
Regarding
"Opportunities, Risks, and Challenges Associated with Growth in Demand for Electric Power in the United States."

May 21, 2024

The leaders of America's electricity producing industries are in an unenviable situation. Like many other professionals, they are engaged in a business that is anchored in making predictions about the future but, more than most, when they get it wrong it's very hard to hide from or casually dismiss the consequences. No one doubts the centrality of electricity to modern life, especially now in our increasingly digital age. The impact of getting forecasts wrong, specifically not having enough electricity available when it's needed, or at prices markets can tolerate, are consequential and widespread.

Forecasts about electricity sit at the center of predicting outcomes at the intersection of three domains: supply, demand and politics. Because electric grids are intended to serve everyone and everything, and do so for decades, traditional utilities and grid operators such as Regional Transmission Organizations, e.g., PJM, ERCOT or the California Independent System Operator (hereafter "utilities"), must evaluate the long-term viability of new or improved technologies for supplying reliable and affordable electricity. At the same time, utilities need to forecast the future demand from greater use of existing technologies and, far more challenging, the emergence of demands from inventions of new electricity-using technologies without being lulled into the complacency of the status-quo. And then of course, arguably more so than for other leaders except those in health-related domains, utilities need to forecast the vicissitudes of future regulations and political forces.

My testimony focuses on the now widely noted, and central issue of this hearing: we appear to be at one of those rare pivots in history witnessing the emergence of significant, and ostensibly unanticipated, new vectors for greater electricity demand. Rising demand not only increases the need for greater capacity to produce electricity but also amplifies the importance of affordability and reliability. The vectors for the new demands are centered around continuing growth in cloud computing now accelerated by the arrival of useful artificial intelligence (AI), the revival and repatriation of electricity-intensive manufacturing (especially for semiconductors), and governmental mandates and subsidies directed at electric vehicles (EVs). Only the first constitutes a technologically net new demand for energy. The other two entail a shifting of demand in terms of sources or locations.

What is clear, as I will summarize below, is that the era slow to flat electric load growth is over and that there is now an urgency to ensuring policies that will ensure adequate, reliable electricity supplies in the coming decade. Talk of an "energy transition" that entails shutting down or replacing conventional power supplies is not only unrealistic but also will have near-term deleterious consequences.

While there is vigorous debate over some of the worrisome uses (and abuses) of AI, from detecting deep fakes to protecting intellectual property, there are also benefits from the new

capabilities that outweigh the challenges. Many such benefits are already clear not least in such things as improving the efficiency of complex supply chains, enhancing security (both civil and military), and accelerating discovery of new therapeutics. It is precisely because of the democratization of AI, a class of software heretofore only available in supercomputers, that one sees such rapid adoption by all manner of businesses. Federal Reserve data shows that private, non-residential investment in information technology (IT) is now running at over $1 trillion annually compared to, for example, $300 billion a year on transportation equipment. In fact, corporate spending on just the AI share of IT investment is now approaching that total spending for transportation equipment.

Estimating the future aggregate impact of shifting on-road energy use from gasoline to electricity for vehicles entails well-understood technical underpinnings that will not change significantly within useful forecasting time periods. The key variable is the level of EV penetration for on-road vehicles wherein an upper bound would be the level of new car sales imagined as feasible by the EPA in its most recent rulemaking. Every $1 billion of new cars in the market leads to, over the cars' operating life, about $200 million of energy purchases. (That total is roughly the same whether for a conventional car or an EV on an equal availability and equal tax basis.) Americans spend over $500 billion a year on new cars.

Similarly, forecasting the range of future demand arising from expanding U.S. manufacturing is anchored primarily in knowing the electric intensity—kilowatt-hours (kWh) per unit of economic output—of manufacturing, especially semiconductor "chips." This is a feature that is well-understood and extremely unlikely to change significantly within the timeframes in which the new facilities may be built. Thus, the key variable is in estimating the number of new chip plants the world will need, and how many of them will be built in the U.S. in the next decade. In both cases, reasonable upper bounds are easy to imagine.

Put in the same economic terms as noted above for cars, every $1 billion spent on new chip plants leads to (an estimated) $300 million in energy (primarily electricity) purchases over a decade. Since the latest forecasts from the semiconductor industry see $250 billion a year in America's annual average spending on new factories, which will in total rival new cars in terms of adding to U.S. energy consumption.

But for the third vector, AI and the broader digital economy, the "cloud," the boundary conditions are far more challenging when it comes to predicting future electricity demand. While it should be obvious that all things digital create net new demands for energy, the challenges in forecasting lie primarily in guessing how much demand will emerge from entirely new kinds of products and services, and the velocity of their adoption. The core challenge is that many of the new technologies are nascent, and most have yet to be invented at all, especially those associated with AI.

Again, in monetary terms, every $1 billion spent on datacenters leads to over $600 million in electricity purchases over an operating decade. Last year, capital spending on datacenters was running at about $100 billion a year in the U.S. Now, the addition of AI-enabled hardware is accelerating both the buildout of datacenters and the energy use per datacenter with at least a doubling in both factors which means, combined, there's a potential four-fold jump in energy use per new dollar of capital deployed in digital domains. That would translate into well over $2 billion in energy purchases over a decade for every $1 billion spent on new AI-infused datacenters.

The AI revolution is on track to add more net new energy demand annually than will manufacturing, or the auto industry, and far more than EVs. And this says nothing about the

spillover effect, the point of using AI in the first place, which is to accelerate economic growth and competitiveness. The arrival of a new way to boost the economy illustrates the long-standing correlation, a veritable iron-law, that links economic growth and rising energy use, especially now electricity.

The boom in both manufacturing of AI-class silicon chips and simultaneously construction of massive AI-infused datacenters should, finally, put to rest the illusion that a digital economy will "decouple" economic growth from rising energy use. But as recently as two years ago, an OECD analysis claimed that "digitalisation can contribute to decoupling economic activity from natural resource use and their environmental impacts." A key question now is whether the current state-of-affairs is a kind of bubble, or whether it signals something more fundamental and if so, just how much more power the information infrastructure will consume, especially now that AI has been accepted as critical to global competitions.

That AI has a voracious energy appetite, that "inference" rather than "calculation" is so energy intensive is not news to the technical community. Consider, for example the results of one recent analysis that showed building, not operating, an AI tool—i.e., the equivalent of energy used to build an aircraft, not fly it—found that a single, modestly small AI application in the "training" phase consumed more energy than driving a Tesla 300,000 miles. Another analysis of building a bigger AI tool, more akin to training ChatGPT, found that application used as much electricity as driving a Tesla four million miles. As with automobiles and aircraft, once built, AI also consumes energy to operate (so-called "inference") which entails as much as a ten-fold greater energy use than training. Meanwhile, the number and nature of potential applications for training and using AI are essentially unlimited.

At a conference several months before the November 2022 unveiling of ChatGPT that ignited the public interest in AI, the CTO of AMD, a world-leading AI chipmaker, observed that current trends pointed to AI, by 2040 consuming a major share of all energy used by the U.S. for all purposes. That won't happen because the AI engines will become much more efficient. While we don't know, yet, is just how fast AI electricity use will grow, we do know that it will grow despite efficiency gains. The underlying energy-efficiency metric for AI hardware has already improved 100-fold in the past half-dozen years, and engineers expect, based on known trends, another 100-fold gain by 2030. But such amazing progress in efficiency won't stem the rise in electricity demand, rather it will stimulate it, the so-called "Jevon's Paradox." And we have seen that play out before. It's what gave us today's global cloud's electricity consumption.

Improving efficiency is a feature of technology progress, and one that is most especially true in digital domains. It was the British economist William Stanley Jevons who first codified this phenomenon in 1865 in a seminal paper focused on the question of whether England would run out of coal given the demands from the industrial revolution. The solution offered by experts at that time was to make coal engines more efficient. Jevons pointed out that improvements in engine efficiency would lead to more, not less, overall coal use. Thus, the so-called paradox. Some modern economists call this the "rebound effect." It's not a rebound; it's the primary purpose of efficiency.

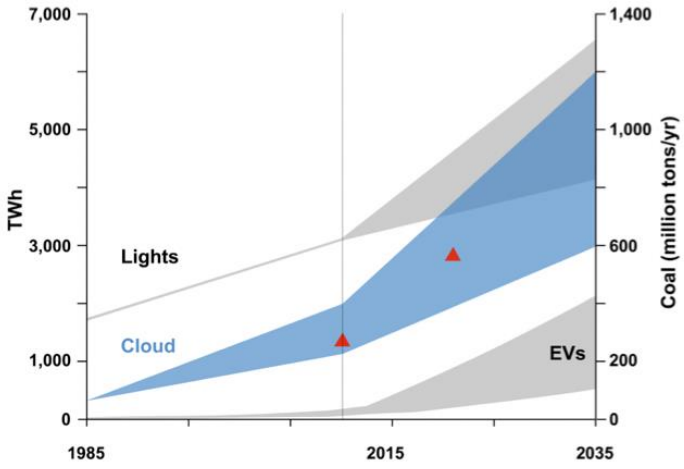
Improved energy efficiency makes it possible for the benefits from a machine or a process to become cheaper and thus available to more people. For nearly all things for all of history, rising demand for the energy-enabled services outstrips the efficiency gains leading to a net gain in consumption. Aircraft today, for example, are three times more energy efficient than the first commercial passenger jets. That efficiency didn't "save" fuel but instead propelled a four-fold rise in aviation energy use.

The same dynamic occurred during our first microprocessor era. Over the past 60 years, the energy efficiency of conventional silicon chips has improved by over *one billion* fold. This means that if a single smartphone today operated at 1980 energy-efficiency, it would require as much power as a Manhattan office building. Similarly, a single datacenter at *circa* 1980 efficiency would require as much power as the entire U.S. grid. But *because* of efficiency gains, the world today has billions of smartphones and thousands of datacenters. That is precisely the efficiency trajectory we will now see for AI hardware and AI-infused datacenters.

Despite the spate of recent headlines evincing surprise at the end of the era of flat growth in electric demand, none of this should have surprised anyone, including the emergence of electricity-hungry AI. Forecasting electricity demand from the first two vectors, more manufacturing and more EVs, is relatively straightforward and can be reasonably bounded. The emergence of the AI, a genuinely new vector for demand, while more difficult to forecast accurately, the technical literature made clear that we should expect significant impact.

I note, for the record, such trends were documented (from the technical literature) in my recent book, *The Cloud Revolution*, published just before the intense media coverage around ChatGPT, as well as in my decade earlier study issued in 2013 (see Figure below), which analyzed the literature to forecast global trends in electricity demand. (Indeed, 25 years ago, my long-time colleague the late Peter Huber and I published articles in both *Forbes* and the *Wall Street Journal* pointing to these realities at the intersection of energy and information.)

Figure
Global Electricity Demand from Three Sources: Forecast Published 2013



Source: *The Cloud Begins with Coal*, August 2013, Mark P. Mills

This forecast, published in 2013, showed the then conventional wisdom (left red triangle) for global cloud electricity use and (at that time) the author's forecast for 2023 demand (upper red triangle). The latter value is now consistent with current industry estimates for actual demand in 2023. The upper bound for EV data is based on the optimistic forecasts for EVs at that time. The actual EV consumption in 2023 was at the lower end of the range illustrated. The study's title derived from the fact that coal had accounted for 50% of net new electricity supplied between 2000 and 2012, contemporaneous with the emergence of net new demand from the cloud.

While the build-out of the modern cloud infrastructure began about three decades ago, we are now at the end of the beginning, not the beginning of the end of that build-out.

The cloud, measured in various terms—the size of the network, the capital deployed, or the energy used—is on track to become the biggest infrastructure ever built by humanity. Global capital spending on (energy-using) hardware to build the cloud and its networks now exceeds global capital spending by all electric utilities on power plants and the power networks.

Similarly, regarding the scale of the global cloud network that connects to people and devices: measured in the combined the length of both physical (cable) and virtual (wireless) connections, that network totals well over a billion miles, vastly exceeding all the world's road-miles.

As for the energy already used by the global cloud infrastructure—operating the networks and datacenters, and building the devices that make it possible—that already rivals the energy used by global aviation, and that's an estimate based on data that are a half-dozen years old. Recent years have seen a dramatic acceleration in datacenter spending on hardware and buildings along with a huge jump in the power density of that hardware, the latter now because of AI being added to that infrastructure.

The question of guessing the extent of electricity demand from AI and the cloud, an emerging technology-industry, is analogous to guessing in the 1950s the future energy demands from aviation. By the early 1950s, the aviation industry was already three decades old and was carrying more passengers that all intercity railways. But the emergence of a new kind of engine, a practical jet engine—equivalent in our time to the arrival of a new class of logic engine, the AI chip now three decades since the cloud began—created the age of modern jet passenger aircraft and global travel. In the two decades that followed the introduction of the Boeing 707, global air passenger-miles increased 15-fold. Today, every $1 billion spent on new aircraft leads to about $2 billion in fuel purchases over the decade of use.

We are in the equivalent of the 1950s when it comes to the expansion of AI and the associated cloud infrastructures.

But whether the benefits from both building and using AI are in fact fully realized in the U.S. will depend on the extent to which the electricity is available, when it's need—in the case of datacenters, that's 7/24—and at prices that are acceptable. It is not too much to imagine that the U.S. could put in place policies that would lead to an AI future that is increasingly offshored much as happened to the aluminum industry. The latter, it bears noting, is also electricity intensive. Producing a $1 billion of aluminum requires using some $400 million in electricity (on low-cost coal-fired Chinese grids). Two decades ago the U.S. was the primary global producer of aluminum, it now has only a few percent share and China, using its low-cost grid produces 60% of the world's aluminum.

Given the emerging scales of electricity demand from the cloud and AI, especially when added to the emerging demands from reshoring manufacturing and promoting EVs, it should be clear that policymakers can no longer entertain the idea of an "energy transition." The nation's electric sector will need full access to all options to ensure enough electricity is produced reliably, and at prices American businesses, and ultimately the public, can afford. <>