USCC TESTIMONY

CHINA'S PURSUIT OF NEXT FRONTIER TECH: COMPUTING, ROBOTICS, AND BIOTECHNOLOGY

PANEL 1: COMPUTING

Addison Snell, CEO, Intersect360 Research
Testimony date: March 16, 2017

EXECUTIVE SUMMARY

For its hearing on "China's Pursuit of Next Frontier Tech," the U.S.-China Economic & Security Review Commission (USCC) is seeking testimony on the current and potential future state of supercomputing innovation worldwide, with an emphasis on China's position on the global stage relative to the U.S. Addison Snell, CEO of Intersect360 Research, provides this written testimony in answer to USCC's questions for the hearing. Mr. Snell will also provide oral testimony to answer additional questions at the hearing on March 16, 2017. Information about Mr. Snell, Intersect360 Research, and the questions asked are in the Appendices of this report. A transcript of the oral testimony will be made available at the USCC's website, www.uscc.gov.

In this statement, we give an overview of the high performance computing (HPC) industry, including analysis of hardware, software, and industry trends. Where relevant, market data from Intersect360 Research is included, particularly for the analysis of significant HPC market segmentations. In the next section, we give a country-level analysis of national supercomputing strategies and programs, for the U.S., China, and other significant countries. In the closing section we give our analysis, conclusions, and recommendations.

While the U.S. still leads by far in the most straightforward market share metrics of production (vendors, supply-side) and consumption (buyers, demand-side), industry indicators show the U.S. is falling behind in the leading edge of advancement. Chinese leadership has apparently recognized the relationship between HPC and economic growth and has set forth on a program to drive the country into a leadership position. The best response to this new challenge is to continue if not increase national support for HPC at all levels.

National supercomputing efforts are essential to motivating investment at the high end. From that point, U.S. companies excel at seizing opportunities to drive markets forward. Against these strengths, the top limitations to Exascale deployments are software and skills. If we do build a system, how will we use it? A key feature of the Exascale Computing Program is its emphasis on co-design, finding end-user stakeholders to collaborate on the design of next-generation supercomputing technologies, bolstered by government funding. We recommend:

- *National initiatives in low-level software tools and programming models, together with stakeholders in industry and academia.*
- *Government-funded partnerships between industry and academia.*
- *Ongoing pursuit of next-generation technologies.*

Regardless of these recommendations, the HPC market will continue, powering new innovations and ideas around the world. Supercomputers today are close to a million times more powerful now than they were 20 years ago. In another 20 years, they could be a million times more powerful still. The leaders in supercomputing will be the ones that do not rest on their achievements, but rather continue to chase the next challenge over each new horizon.

STATE OF THE HIGH PERFORMANCE COMPUTING INDUSTRY

High performance computing (HPC) is the use of computational servers and supercomputers, together with their attendant storage, networking, and software components, for the modeling and simulation of complex phenomena. As such, HPC is a critical segment of the information technology (IT) sector, powering innovation and scientific discovery in a wide range of disciplines across industry, public-sector research, and national defense. From clusters of a few servers to the world's largest supercomputers, HPC enables the exploration of new frontiers in science and engineering, by complementing or even replacing physical testing and experimentation that otherwise may be expensive, time consuming, or even impossible.

The Importance of HPC

For centuries, scientific method has been based on the two pillars of theory and experimentation. Today there are many who call out computational modeling and simulation as a third fundamental pillar of science, as more discoveries are made possible by HPC. As such, HPC is linked inextricably to scientific research, yes, but also to industries such as manufacturing, oil exploration, pharmaceuticals, chemical engineering, and finance, as well as to national defense, homeland security, and cybersecurity.

HPC in Research

The largest supercomputers in the world are concentrated in government and academic research centers, addressing "grand challenge problems" such as predicting the path of a hurricane, modeling the spread of cancer in the body, or simulating the neural pathways of a human brain. But it is not only in these world-renowned supercomputing centers that discovery takes place. Researchers at almost any center of higher learning will leverage HPC in some dimension, be it astrophysics, biochemistry, geology, or archeology, whether on their own lab systems or by accessing resources available via public cloud computing.

HPC in Industry

Although the public perception of supercomputing is predominantly linked to public-sector research, more than half of HPC spending worldwide comes from commercial organizations. Scientific innovation begets advancements in engineering and technology, and industrial organizations seek competitive advantage from investment in computing. Some of the top industries that leverage HPC are:

- ***Manufacturing:*** Digital product simulation allows manufacturing companies to design, test, and modify virtual products in a computer environment without the time and expense of building and testing physical prototypes. At a large scale, this applies to car manufacturers conducting virtual crash tests of automobiles and airplane manufacturers modeling the aerodynamics and noise of wind over a wing. At a smaller scale, increasing numbers of consumer products manufacturers are deploying HPC for any number of household products—for example, designing a bottle such that it uses less plastic yet is still less likely to break if dropped from table height.
- ***Energy:*** Oil and gas exploration companies use HPC to model the earth's crust from seismic readings. Doing so lets them find the underground, prehistoric lakes and riverbeds they can extract oil from today. As oil and gas are extracted, the companies continue to run simulations to maximize yield. HPC has even helped oil companies return to fields they had previously abandoned, armed with new technology and new information. Oil companies tend to use some of the largest, most powerful supercomputers in the commercial sector.
- ***Pharmaceuticals:*** Drug discovery is an expensive R&D proposition. For every 100 targets, pharma companies hope one will prove viable to come to market. HPC helps those companies "fail faster," eliminating losing designs earlier in the process, so that more time and effort is spent on the most promising candidates. It was HPC that enabled the burgeoning field of genomics, and HPC has

applications across computational chemistry, molecular modeling, pharmacokinetics and pharmacodynamics (PK/PD), bioengineering, and biostatistics.

- ***Chemical engineering:*** Countless new products and innovations begin with the introduction of new plastics and polymers. Chemical engineering companies use HPC in their manufacturing process, for example, by using modeling to optimize their mix tanks for improved yield. In the future, HPC may be increasingly used in the design of these chemicals as well, similar to how HPC is used in other industries.
- ***Electronics design:*** As semiconductor designs continue to shrink, it becomes perpetually more costly and difficult to design and test them. Semiconductor companies use HPC for electronic design automation (EDA) of their chips, reducing costs and improving yields.
- ***Finance:*** HPC has revolutionized how the finance industry works. Banks use large HPC clusters for econometric simulations, running new products through millions of potential worldwide economic situations before selling them to the public. In addition to risk management, financial services companies deploy HPC for high-frequency trading, pricing, and a wide range of analytics applications, such as fraud detection. In all, finance is one of the largest consumers of HPC technology in industry (just behind manufacturing, if large product manufacturing and consumer product manufacturing are combined), and also one of the fastest growing.

HPC in Defense

With how intrinsic HPC is to the advancement of scientific research and industry, it is no surprise that HPC is also critical to national security. For starters, advancements in industry and research can have direct translation to military efforts, for example, in the manufacture of a new helicopter or the availability of an ultra-local weather forecast supporting an operation. Many supercomputing endeavors may be classified when it comes to military and defense issues, but there nevertheless several categories of applications we can be sure of:

- ***Combat simulations:*** One of the most straightforward ways in which supercomputing is deployed for national defense is in combat simulations, ranging from “person-in-the-loop” models like flight and tank simulators to large-scale wargames. (The 1983 movie *WarGames* was based on exactly this concept.)
- ***Nuclear stockpile stewardship:*** HPC allows the maintenance of a nuclear stockpile without the need for nuclear testing. Nuclear explosions can be simulated using supercomputing technologies.
- ***Defense and response simulations:*** In addition to simulating combat, HPC can be used to simulate defensive responses to possible attacks, modeling for example, how to respond to an attack on the electrical grid or the release of a chemical agent in a metropolitan area. (Theoretically, these simulations could also be used offensively, not just defensively.)
- ***Intelligence and security:*** HPC is used in the gathering and analysis of data streams that can provide evidence of planned attacks or point toward an enemy’s weaknesses. Information analysis can also be used to make borders safer, including linking the analytics with emerging advancements such as facial recognition.
- ***Cybersecurity and cyberattack:*** Cybersecurity is an issue that affects companies and private citizens, but at the government level, it is a matter of national security. HPC is now being used in some cases to model defensive responses to cyberattacks. That said, this is an instance in which HPC is used offensively more frequently than defensively, and at larger scale. Supercomputing can be used to break passwords, to hack into systems, and to analyze any information that is acquired.

What Is an HPC System?: Eras of Innovation

Like in the industries fueled by HPC, there is a rapid pace of change in HPC itself. The very nature of HPC is that there is always a harder problem to solve. Simulations take on grander scale and more realism. Models get built with higher fidelity and more degrees of freedom. Until one day we reach the end of science and declare there is nothing left to be discovered or invented, there will be a need for a more powerful computer to enable insight and innovation.

When we examine the HPC industry, we consider what constitutes HPC at the low end of the market. At a minimum, we consider scientific and engineering tasks that can be run in parallel over at least two “nodes”: otherwise independent elements that are being used in concert to address the task. Technical computing extends down to applications that can be run on individual PCs or workstations, but we consider this to be below the threshold of the HPC market.

As the HPC market has grown and spread, the systems themselves have gone through transformations in architecture. The ongoing evolution in how an HPC system is built has continually changed the dynamics of the industry in terms of who has access to HPC systems, how difficult they are to use, and how much they cost.

While innovation in HPC is continuous, a long-view history of the industry can segment development into distinct eras:

- ***Specialized design (technologically contained / architecturally distinct):*** Originally, supercomputers were relatively self-contained devices. A single technology vendor would design all the pertinent hardware and software components to build a scalable system, including processing units, networks, and operating systems. These designs were inherently custom to each specific vendor, and there were only a few companies that made them, such as CDC, Cray, IBM, and NEC. Over time, enterprise computing engendered an evolution in which server technology—using RISC processors and variations of the UNIX operating system, from companies such as IBM, HP, Sun Microsystems, Silicon Graphics, and Digital Equipment Corp.—could be scaled up for HPC. These components and operating environments were more familiar to more users, but the systems were still unique to their specific vendors.
- ***Commodity clusters (technologically disaggregated / architecturally identical):*** In the 1990s, some HPC users—including major supercomputing centers—began constructing HPC systems as “clusters” of lower-cost, industry-standard servers, connected over an Ethernet network. This allowed users to build systems to any size while getting components from any vendor using x86-architecture processors from Intel or AMD, running versions of the open-source Linux operating system. (Linux is still the dominant operating system for HPC.) However, because the clusters were in fact comprised of distinct components, existing software modules had to be rewritten for the new architecture. Many people argued that the cluster model was less efficient and harder to program for, but after 10 years, clusters had become the dominant paradigm for building HPC systems. Over time, many enhancements became available (such as 64-bit processors, blade form factors, or faster networks), but the cluster architecture and its accompanying software model remained.
- ***Specialized elements (technologically disaggregated / architecturally distinct):*** Today we are seeing the HPC industry in the early stages of another architecture transition[i], the seeds of which were sown about 10 years ago, in a pair of parallel developments. For one, microprocessor speeds (measured in GigaHertz, GHz) plateaued, and microprocessor companies like Intel and AMD focused instead on putting multiple processing “cores” onto a single chip. For systems that serve multiple applications, each of which only requires a fraction of a core (as in a typical PC environment), this makes little difference, but for an HPC application that already must span multiple processors, the change required a new level of optimization in software. Simultaneously, other processing options began to come to the fore, including increasing interest in various types of “accelerators,” additional processing elements that can be added to a system in order to boost the performance of certain functions. The biggest push in acceleration came from NVIDIA, a gaming company that found a market for its graphics processing unit (GPU) products in accelerating HPC applications.

With these and multiple other developments, there are now myriad options for building HPC systems out of specialized processing elements. And while this sounds promising in terms of providing options, it also presents a tremendous challenge in determining (a) which architectural components will best suit a given problem and (b) how to rewrite software in order to run most efficiently. Furthermore, applications have varying requirements. As a result, 88% of HPC users now say they expect to support multiple architectures over the next few years, attempting to align application workloads to the elements that best suit them.[ii]

This is important in that we are now at another crossroads in the HPC industry, at which HPC-using organizations must revisit how software components must be built to take advantage of new technology developments, against a backdrop of uncertainty as to which technologies will win out in the market. As we move into the new generation of supercomputing, there could be significant change in where leadership in the industry comes from.

Measuring Supercomputing

The most common metric of supercomputing performance is “flops,” short for floating-point operations per second.[iii] (In computer parlance, a floating-point operation is one with an arbitrary number of decimal places, such as “2.4 x 1.624 = 3.8976”; this is in contrast to integer-only operations.) As with other computer terms like “bytes,” flops are usually expressed with a numerical prefix: “100 Megaflops” means 100 million calculations per second.

Of course, the HPC industry is well beyond Megaflops. At the uppermost echelon of supercomputing, we have also transcended Gigaflops (billions of calculations per second) and Teraflops (trillions), into the range of Petaflops (quadrillions, or millions of billions). Currently, the top supercomputer in the world, measured by theoretical peak performance, is the Sunway TaihuLight[iv] at the National Supercomputing Center in Wuxi, China, with a peak of 125 Petaflops—a mere factor of eight short of the next prefix marker, Exaflops (quintillions, which can be thought of as billion-billions or million-million-millions).

In the previous paragraph, “theoretical peak” is worthy of emphasis, because actual delivered performance will fall somewhere short of this mark, depending on the software application and its programming. For this reason, it is common to look at benchmarked supercomputing performance for specific applications.

The TOP500 List [v]

The semi-annual TOP500 list is a ranking of the top 500 supercomputers in the world according to the LINPACK benchmark[vi], a straightforward exercise of solving a dense matrix of equations. On this benchmark, the Sunway TaihuLight is still top, with a delivered performance of 93 Petaflops, or about 74% of its theoretical peak. The second-ranked system is also in China, the Tienhe-2 (or “Milky Way 2”) system at the National Supercomputer Center in Guangzhou, with a delivered LINPACK benchmark performance of 34 Petaflops, against a theoretical peak of 55 Petaflops (62% efficient).

The next three systems on the current list (published November 2016) are all in the U.S., at labs run by the Department of Energy: “Titan” at Oak Ridge National Laboratory, 18 Petaflops LINPACK, 27 Petaflops theoretical peak (65% efficient); “Sequoia” at Lawrence Livermore National Laboratory, 17 Petaflops LINPACK, 20 Petaflops theoretical peak (85% efficient); and “Cori” at Lawrence Berkeley National Laboratory, 14 Petaflops LINPACK, 28 Petaflops peak (50% efficient).

The next five systems in the current ranking include two in Japan, one in Switzerland, and two more in the U.S. The rest of the top 25 also include systems in the U.K., Italy, Germany, Saudi Arabia, and France. Among the rest of the top 100, there are systems in South Korea, Poland, Russia, the Czech Republic, Switzerland, Sweden, Finland, Australia, and the Netherlands.

The TOP500 list is frequently used as a proxy for the state of the HPC market, but this has several drawbacks. First, a system need not be represented on the TOP500 list. Because running the LINPACK benchmark itself requires the dedication of time, expertise, and resources, inclusion on the list is usually driven by a desire to be recognized. Many commercial organizations decline to participate in TOP500 ranking. (The top-ranked commercial system on the current list is at #16, a 5 Petaflops (LINPACK) system owned by Total Exploration Production, an oil exploration company in France.) Even in the public sector, there are several systems not listed; perhaps most notable among these is the Blue Waters supercomputer at the National Center for Supercomputing Applications at the University of Illinois, which may in fact be the most powerful supercomputer at an academic site in the U.S.

Participation in TOP500 is also often driven by a campaign of inclusion, particularly by system vendors who wish to claim a strong position for marketing benefits. Not only may vendors encourage their customers to participate, but the submission rules also allow systems to be included by proxy: If the LINPACK benchmark is run and verified on one system, then it can be assumed to run at least as well on any system of the same configuration with at least as many processing elements. Some of these submissions' locations are identified only generally by industry, such as "Government" or "Automotive." This is more common in the lower tiers of the list, where the systems are not as distinctive.

In 2016, the proportion of systems on the TOP500 list based in China suddenly expanded. This does not mean that China suddenly discovered supercomputing, but rather, that (in a sense) China suddenly discovered the TOP500 list, and that there was a somewhat nationalistic desire for its systems to be counted.

The TOP500 list is also often criticized because the LINPACK benchmark used for the ranking is not representative of general mixed supercomputing workloads. In fact, it is a relatively straightforward benchmark. To use an athletic metaphor, it measures sprints, not decathlons, and it certainly does not extend to also include crossword puzzles and a talent contest. As such, the benchmark taxes the computing elements of a system far more than others, such as the memory, the network, the data storage, or (quite pointedly) the programmability.

One might own a supercomputer that in its time excels at LINPACK and little else. These are often derided as "stunt" machines, good for PR value but not for science. However, it is important to acknowledge that we have heard this debate before. When clustered architectures became the new supercomputing paradigm, there was little software available to run on them, and programmability was viewed as a major hurdle. The first machine to achieve a Teraflop on LINPACK was the ASCI Red cluster, first installed in 1996 at Sandia National Laboratories, as part of the Accelerated Strategic Computing Initiative (ASCI) under the DOE's National Nuclear Security Administration (NNSA). Although ASCI Red did run its intended applications in nuclear stockpile stewardship quite well, it became a touchstone of debate in the supercomputing community, in that it was not at the time viewed to be a "general purpose" system for scientific discovery. In the years that followed, clusters came to dominate on LINPACK and the TOP500 list, but it took years of effort to get many scientific applications to use clusters efficiently. For some applications, non-clustered (single system) supercomputers are still preferred today.

A final criticism of the TOP500 list is that it is not representative of the HPC market as a whole, which is far broader than 500 (mostly public-sector) systems at the top end of the market. A thorough industry analysis looks quite different than an analysis of the TOP500 list, in many segmentations.

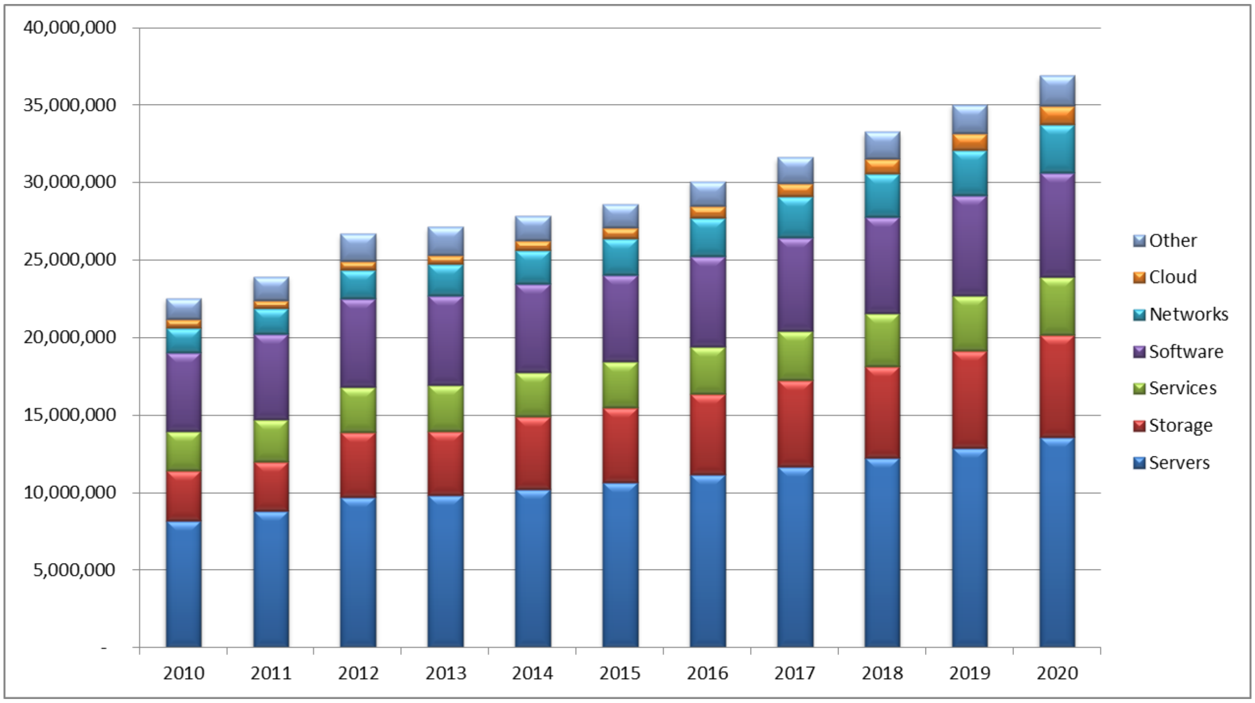
Nevertheless, the TOP500 list is a useful tool, and the best one available for ranking individual supercomputing systems at the top of the market. It provides a starting point with interesting data, and it helps to prod investment. But careful, thoughtful, thorough analysis of the supercomputing market must go beyond the TOP500 list and the LINPACK benchmark.

Current HPC Market Dynamics

In 2016, the total worldwide HPC industry was a $30.1 billion market[vii], including all product and services categories, such as servers, storage, software, networking, services, and even HPC conducted over public cloud resources. This figure does not include internally budgeted items, such as power consumption or personnel costs, which can also be considered as part of a total HPC budget for an organization.

Of this $30.1 billion, the largest component is servers, which include all types of HPC systems. At $11.1 billion in revenue, servers comprise 37% of the HPC market. The next largest components are software ($5.8 billion, 19.4%) and storage ($5.2 billion, 17.3%). Although networking appears as a relatively small component ($2.5 billion, 8.2%), it is important to realize that system networking is most often included with an HPC cluster, and therefore is absorbed into the cost of servers.

Figure 1: Total Worldwide HPC Market, 2010-2020, ($000)[viii]
2010-2015 actuals, 2016-2020 Forecast



HPC in the Cloud

Cloud computing ($774 million in 2016, 2.6%) is a small portion of the HPC market. Here we refer to money spent by HPC end users to run their applications on a public cloud resource (such as Amazon Web Services, Google Cloud, Microsoft Azure, or SoftLayer). We are not counting the entire infrastructure owned by cloud service providers as part of HPC; for a discussion of that market, see “The Hyperscale Market,” below.

The utility computing model—renting infrastructure and paying for it on an as-used basis—has been part of the HPC market for decades, but always as a very small component. Cloud computing has made this easier, and many organizations are choosing to outsource large portions of their enterprise infrastructures to cloud providers. Indeed, beginning in 2014, we began to see an uptick in the usage of cloud for HPC, and we project a high growth rate throughout our forecast period.

However, cloud will continue to be a small part of the total HPC market. Cloud computing makes the most economic sense for those who need to fully utilize a resource for a short amount of time, much like renting a car or taking a taxi. Those who need to use the resource all the time often find it more economical to buy it than to rent it.

Within HPC, public cloud resources are most often used by academic researchers, who may be able to utilize cloud in burst for their projects or dissertations. In industry, the most significant usage currently comes from pharmaceutical companies, some of whom are reducing their internal HPC footprints and using cloud resources for occasional peak workloads.

Public cloud resources do make HPC more accessible to a wider community. However, the vast majority of serious HPC users will continue to own on-premise systems.

Sunway TaihuLight and Other “In-House” Systems

As noted above, the Sunway TaihuLight system is currently the most powerful supercomputer in the world, according all available public metrics. Despite this fact, it is difficult to include in HPC industry models by revenue, because the system was *built, not bought*.

The Sunway TaihuLight supercomputer at Wuxi is the product of a Chinese national initiative to build a supercomputer from emerging domestic technologies. Little information is publicly available, but a total budget of $300 million was announced. If we took that figure as a portion of the 2016 HPC market, it would represent 1% of total spending. However, this would be misleading. Most of this money was spent on manpower to design technologies. The personnel costs of system vendors aren’t considered to be market spending, and this is also the case for this system at Wuxi. Only a minority of the $300 million was spent on components from established technology vendors. (For example, some silicon parts were purchased from networking company Mellanox, but the National Supercomputer Center says it designed its own system network based on these parts.) The true retail value of this system is therefore unknown and is not part of the revenue reporting.

At a smaller scale, any organization might build an “in-house” system. To the extent that components (such as servers or networking gear) are bought from established vendors, this revenue is accounted for in the market model. But if an HPC system is truly invented or assembled from materials lying about, there is no revenue to be tracked.

HPC Market Segmentations

As described above, HPC systems range from small clusters of only a few nodes to the world’s largest supercomputers, with use cases spanning industry, academia, and government worldwide. Here are most of the most pertinent top-level market segmentations.

Figure 2: Vertical Market Distribution of HPC Revenue[ix]

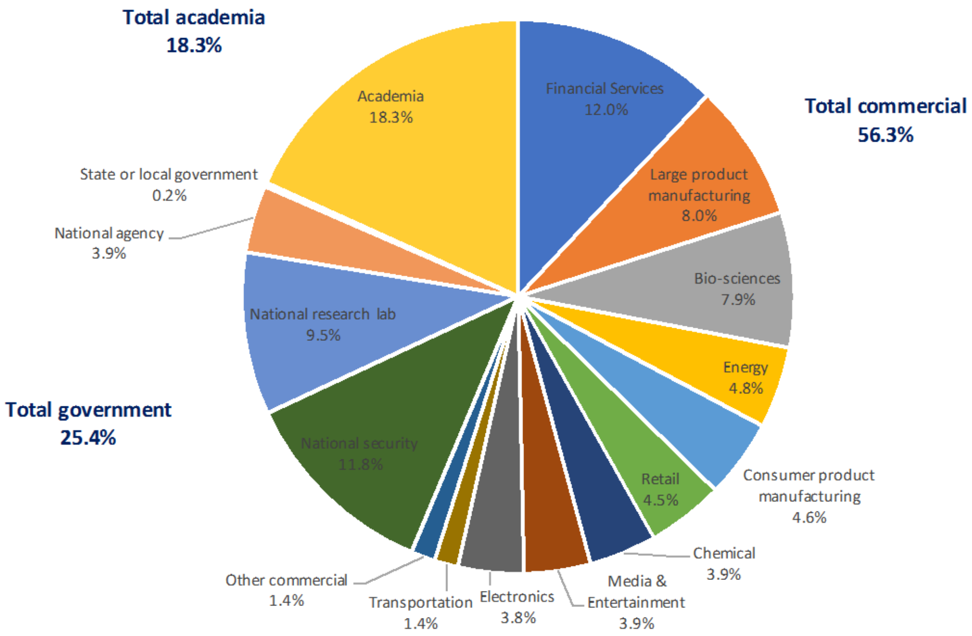


Figure 3: HPC Server Revenue, by Product Class[x]

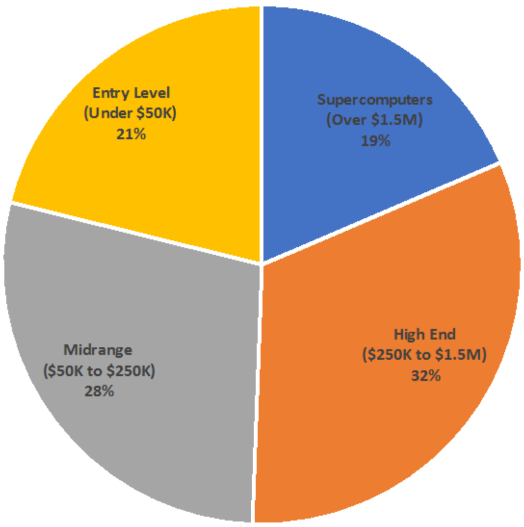
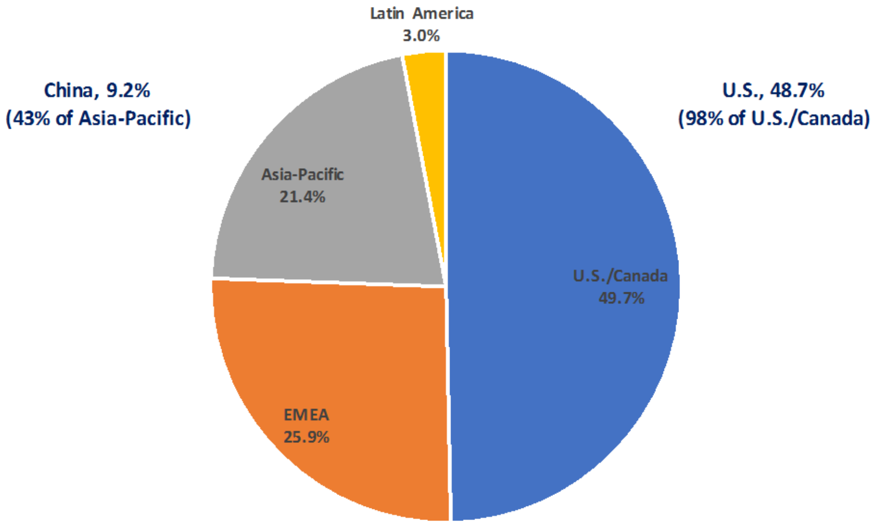


Figure 4: Worldwide HPC Revenue, Geographical Distribution[xi]



Although the world's two most powerful supercomputers are in China, and China also has more systems on the TOP500 list than any other country, there is in fact five times more HPC usage (by revenue) in the U.S. Even if we added $1 billion (all in one year) to the Chinese HPC market to represent an arbitrary high value for Sunway TaihuLight, the overall picture would not be much different.

The reason for this is twofold. First, China has done a remarkable job at a national level of painting a picture of a stronger market, versus no such organized effort in the U.S. to make the market seem larger. Second, and more importantly, the U.S. has a much more robust breadth of HPC usage across multiple industries. Compared to the U.S., China is still a relatively poor country, and there is less penetration of HPC through multiple markets.

That said, it does leave the opportunity for ongoing strong growth in HPC in China. Furthermore, China's investments in new supercomputing architectures, using domestic technologies, should not be discounted, particularly in light of the architectural shift now in play in the HPC industry.

Key Technology Component Vendors

As discussed above, HPC environments tend to be technologically disaggregated. That is, the key components of an environment often come from multiple different sources. The following are some of the key vendors from across the HPC landscape.

- ***Servers:*** Worldwide, the overall market share leader in HPC server revenue is Hewlett Packard Enterprise (HPE). HPE recently acquired SGI, a onetime big name in HPC that still retained a small but loyal following and some unique technology. Dell EMC is not far behind HPE in market share. IBM used to be the biggest vendor, until it sold off its Intel-based server lines to Lenovo. IBM still retains some share with its systems based on IBM POWER processors. Cray (which specializes in the high-end supercomputer segment) and Penguin Computing are also noteworthy U.S.-based vendors. While

Lenovo is based in China, it is a worldwide brand and now a tier-one vendor for HPC servers. Although it sells predominantly in China, Inspur also has significant share. Huawei has recently emerged as an up-and-coming brand. Supermicro sells some systems under its own brand, and many more through OEM partnerships. Atos/Bull is also a serious supercomputing provider based in France, as is Fujitsu in Japan.

- ***Processing elements:*** Intel is the giant among the processor vendors. Over 90% of the HPC market is built on Intel-based servers. IBM retains some share with its POWER processors, and AMD is still a competitor as well. However, other elements besides the main microprocessor are now at play, with various types of accelerators and co-processors. The most significant of these are graphics processing units (GPUs) from NVIDIA. Another category of co-processors is field programmable gate arrays (FPGAs). Here Intel is also a player, thanks to its acquisition of Altera. Xilinx is the most notable competitor.

  Another alternative that has become a touchpoint of research is ARM processors, the comparatively lightweight, lower-power processing components common to mobile devices such as iPhones and iPads. Companies such as Cavium, Qualcomm, and AppliedMicro have come out with commercial, 64-bit ARM varieties, and several supercomputing initiatives have been based on the concept. ARM offers open licensing, allowing multiple vendors and research centers to design around the processors. Fujitsu has announced its intentions to use ARM processors in its “Post-K” Exascale systems[xii], and China has its own efforts to design 64-bit ARM processors domestically. ARM Holdings, a U.K.-based company, was acquired by the Japanese conglomerate SoftBank in 2016.

  Many other companies provide unique solutions. To reduce the reliance on outside technology, the Chinese government has invested in the development of its own domestic processing technologies.
- ***Networking:*** Over the past ten years, the dominant high-end system interconnect option has been a technology called InfiniBand, which is now offered almost exclusively by Mellanox, a high-performance networking specialist with dual headquarters in the U.S. and Israel. Intel has the ability to sell InfiniBand (based on its previous technology acquisition from QLogic), but Intel primarily is pushing its own high-end interconnect, OmniPath, which is relatively new to the market. Though it has no intentional HPC strategy, Cisco sells enough Ethernet solutions to be counted among the market leaders. Some of the supercomputing vendors, such as Cray, Bull, and Fujitsu, also have their own custom networking elements to support their large systems.
- ***Storage:*** Storage in HPC has no dominant vendor. Dell EMC has the greatest share, followed by NetApp. Other significant players include Seagate, Hitachi Data Systems, HPE, and IBM, along with HPC niche specialists DDN and Panasas. It is worth mentioning that while this testimony focuses primarily on computational elements, for many applications and workloads, the configuration of high-performance storage for scalable data management can be the more significant investment.

The Hyperscale Market

Not included in these HPC market models are the “hyperscale” infrastructures at large internet application and cloud providers. This includes companies such as Google, Facebook, Amazon, Microsoft, AT&T, Apple, eBay, Baidu, Tencent, and Alibaba. These companies do operate at scale, and they do consume a certain amount of high-performance technologies.[xiii]

Beginning in 2007, Intersect360 Research maintained “ultrascale internet” as a segment of the HPC market. But by 2014, it was evident that what became better known as “hyperscale” had grown and evolved into its own segment, with distinct dynamics from the HPC market. We define hyperscale as ***arbitrarily scalable, web-facing application infrastructures that are distinct from general enterprise IT***. In the top tier of the market, there are eight companies that spend over $1 billion per year on hyperscale infrastructure. There are dozens that spend at least $100 million per year, and hundreds more that spend over $1 million. In all, the hyperscale market represented $35 billion of the enterprise IT market in 2016, concentrated into the high-end providers. This is separate from the $30 billion HPC market.

The hyperscale market is significant to this discussion for several reasons. First, it provides a second, high-growth IT segment for the consumption of high-performance components. This is a powerful attractant for technology providers. Selling to just one of these top-tier hyperscale companies can mean tremendous success. Second, techniques and technologies from the hyperscale market can migrate into HPC, whether they are innovations in hardware, in software, or in the facilities management challenges of operating large systems at scale. And third, the hyperscale market has been the epicenter of development for the rapidly developing field of artificial intelligence (AI), fueled by new capabilities in the domain of "machine learning." (See section on Artificial Intelligence in "Looking Forward," below.)

Despite all these similarities, there are key criteria that make hyperscale different from HPC. The most important has to do with function. In serving up internet applications (whether in search, social media, content streaming, retail, or any other hyperscale domain), hyperscale infrastructures are designed to handle large numbers of small jobs. By contrast, HPC infrastructures are designed to handle comparatively small numbers of larger jobs. As such, most hyperscale infrastructures would be poor fits for HPC workloads. The difference is subtly implied in the names: In hyperscale, the greater emphasis is on scale; in High Performance Computing, the greater emphasis is on performance.

Big Data and Analytics

Another key proximity area to HPC has been the field of big data and analytics. Big data created a new category of enterprise IT applications that has some of the same characteristics of HPC applications. Analytics seeks out answers in data, with metrics like *time to insight* that measure the effectiveness of big data programs. As such, metrics of performance become important criteria in evaluating technologies for big data. Surveys in 2012 and 2013 confirmed that "I/O performance," the ability to move data into and out of a system, was the greatest "satisfaction gap" in technologies for big data—the area with the greatest unmet need relative to satisfaction with existing solutions.[xiv]

However, big data has not proved to be a major boon for the HPC market. Although many organizations launched big data initiatives, for most organizations, their spending patterns did not change much. Companies began with the data, the storage, the servers, the networks, and the personnel they already had, and they did the best they could. (They may have invested in new software, but this was often inexpensive or even free. Services spending did get a small boost.)

The most significant technology that did emerge from IT's big data awakening was flash storage, or solid state disks (SSDs). These devices offer faster I/O performance than traditional spinning hard drives, albeit at usually a higher cost per byte of storage. Many organizations inside and outside HPC now incorporate SSDs as part of their total storage hierarchy.

Analytics is a category of application that can be run on any system, whether or not it is HPC. "Analytics" is also a very broad term that can be stretched to cover many types of computing. Some HPC systems run big data workloads, but not all big data workloads are HPC.

The Race to Exascale

As mentioned above, there is an insatiable need for ever more powerful HPC systems to solve increasingly complex scientific problems. The TOP500 list provides a measuring stick for achievement, as new deployments leapfrog over their predecessors in how many flops they deliver. This pursuit takes on extra significance as we approach the thousand-fold increases that usher in a new prefix era.

In 1996 ASCI Red (discussed above in the section, “The TOP500 list”), was the first supercomputer to achieve 1 Teraflop of performance, perceived as a great victory for the cluster experiment. Twelve years later in 2008, the Roadrunner system, operated by NNSA at Los Alamos National Laboratory, achieved the milestone a thousand-fold greater, becoming the first system to deliver 1 Petaflop of performance.

We are approaching the time when the first Exascale system will be deployed. *Exa-* is the next prefix on the list; 1 Exaflop = 1,000 Petaflops. An Exaflop is one quintillion calculations per second; that’s one billion-billion, or one million-million-million, or if you prefer, 10^{18} or 1,000,000,000,000,000,000. If those speeds seem absurd, rest assured that there are applications that will demand them, not only in scientific research, but in industry as well.

The following is an excerpt from a 2014 report by the U.S. Council on Competitiveness, generated by Intersect360 Research, on the benefits of supercomputing for U.S. industry.[xv] The first key finding of this report was that “U.S. industry representatives are confident that their organizations could consume up to 1,000-fold increases in computing capability and capacity in a relatively short amount of time.”

> There is tremendous optimism across industry that increases in capacity would be consumed. Looking at their most demanding HPC applications today, 68% of respondents felt they could utilize a 10x increase in performance over the next five years. Perhaps more surprisingly, 57% of respondents – more than half – say they could make use of a 100x improvement in performance over five years, and 37% – more than one-third – still agreed when the threshold was increased to 1,000x. This finding is supported by the qualitative interviews, as follows:
>
>> *There are two Holy Grails at exascale that I am just dying for. One of them is computational steering, taking the engineer, the scientist, the doctor, the accountant, putting them in the chair, give them the joystick, basically running through the application and continuously optimizing to whatever state they want to be at. The other Holy Grail is being able to do digital holography, where I can truly create virtual objects, which is the ultimate VR [virtual reality]. … To me, that unlocks human creativity, and we have another Renaissance period, a scientific Renaissance.* [Interview 12, Consulting services for industrial supercomputing adoption]
>>
>> *I don’t think anybody can exactly articulate the extent to which it’s going to change how we do product development, but it could be radical. It could take our development time to less than half, perhaps, if you don’t halve to build prototypes and have systems in the field and do all of the testing virtually. Who knows?* [Interview 3, Large product manufacturing]
>>
>> *In a research mode, we can evaluate a [single] design, but to put it into full production and try to evaluate the [entire] product line, it’s impossible at that level. We can impact things at a research level – to try to understand the benefit, can we go in this direction – but to really have a broad impact on the whole product group, it’s prohibitive. We’re going to need somewhere between 10x and 100x in order to achieve that.* [Interview 7, Large product manufacturing]
>
> One leader from the financial industry did provide a detailed roadmap of what his organization could do with each new level of application scalability:
>
>> *There’s a whole hierarchy that happens in every product in finance. When people start trading a product, the first thing they need is a price. They need to be able to compute an arbitrage-free price based on other securities. … That involves having a model that you can calibrate to the market and price the*

> *security. That's one level of computation. If it's a complicated model, it can take significant computing power to do it.*
>
> *Now, the next level up, once you can do that, you want to say, how is the price going to change if the market changes? Now you have to perturb all the market input models, and there could be five or 10 or 20 or 30, and re-compute, so now you're talking about increasing the level of computation you need by an order of magnitude.*
>
> *And then once you can do that, there's two other directions it goes. Now I want to analyze the strategy that's involving the security, so I want to pull historical data and try running out the strategy using this model every day over the last five years. So now you have a huge amount of computation to run each of these tests, another couple orders of magnitude. And then once you're trading these successfully you have a portfolio of them that you need to analyze how the whole portfolios going to behave, so it's another several orders of magnitude.*
>
> *As the computing gets faster it makes more things possible. … Once your computing catches up and you can do it on an interactive basis, you can respond to market changes, and it opens up a whole new world. When you have to do your portfolio analytics overnight, then it's a different world than when you can do them in real time, interactively, where I can say, 'Oh, the market moved suddenly. How does that impact my entire portfolio? Can I track my VaR [value at risk] as the market moves?' That's an innovation that could have a major impact on the markets.* [Interview 2: Financial services]

Exa-scale vs. Exa-flops

While the pursuit of greater levels of performance has been undiminished, one subtle detail has changed. In the industry vernacular, it is common to discuss Exa*scale*, not just Exa*flops*. The word *Exascale* does not have a specific meaning, but its usage is born from the discussion of the usefulness of a supercomputer. What good is an Exaflop if it cannot be attained by any real scientific application? In that vein, *Exascale* can be thought of to mean "having real applications that run at an Exaflop." (In practice, however, many people do not make a distinction, and once any Exaflop supercomputer is built, it is likely that many will proclaim, "Exascale has arrived!")

If it seems inevitable that such a system will be built, and soon, there are nevertheless dramatic challenges to be overcome. Some of the most significant are:

- ***Power consumption:*** The Sunway TaihuLight system consumes over 15 Megawatts of power for its 93 Petaflops of LINPACK performance. That ratio of 6 Teraflops per Kilowatt is already second-best among the top ten supercomputers in the ranking. (The best in the top ten is the eighth-ranked Piz Daint system at the Swiss National Computing Center, CSCS, which delivers 9.8 Petaflops for 1.3 Megawatts, a ratio of 7.5 Teraflops per Kilowatt.) Even at 10 Teraflops per Kilowatt, an Exaflop system would require a power budget of 100 Megawatts. The U.S. Exascale Computing Project sets a goal of delivering an Exaflop system with a power budget of 20 Megawatts, a ratio of 50 Teraflops per Kilowatt.
- ***Reliability:*** The more components you add to a system, the greater the odds that one of them will fail. The Sunway Taihulight system contains over 10 million processor cores. If only one in a million fails on a given day, there are 10 failures per day, and that does not take into consideration failures in memory, data storage, or networking. An Exascale system may have an order of magnitude more components than that. Systems of this scale will need to handle individual component failures gracefully, without significantly degrading the system as a whole.
- ***Programming:*** Considering the underlying hardware changes at play, as well as the increasing diversity and specialization of technologies, this may be the greatest challenge of all. Basic algorithms and programming models need to be revisited for this level of scale, and what works best on one type of supercomputer may not work efficiently (or at all) on another.

Who Will Get There First, and When?

The Chinese have a substantial edge right now at the zenith of the supercomputing market, and a funded plan to drive to Exascale. There is a good chance that China will deploy an Exascale system (built “in-house”) by the end of 2019. Japan previously had an Exascale plan on a similar timeframe, but recent delays mean Japan likely won’t achieve Exascale until 2020 or 2021 (likely to be built by Fujitsu).

The U.S. had initially planned to deploy its first two Exascale systems in 2022 to 2023, about three years after the Chinese, with one system based on Intel architecture, and another by IBM with NVIDIA GPUs. In November 2016, the U.S. Exascale Computing Project ratified a strategy to introduce an additional “novel architecture” sooner, by the end of 2021, a full year or more ahead of the originally planned systems.[xvi] [xvii] In that timeframe, the U.S. would deploy closer to China and Japan. France and Russia could also field Exascale systems in a similar timeframe.

Looking Forward

Looking ahead, there are some technologies and applications that have the potential to provide a discontinuous advancement in the way that HPC is done.

Machine Learning / Artificial Intelligence

Hyperscale companies are using “machine learning” techniques to make significant advancements in artificial intelligence. Deep learning involves two steps: training and inference. In the training step, massive amounts of data are used to analyze patterns. For example, there may be millions of photos that are tagged to indicate “cat,” and millions more similarly tagged, “not a cat.” The training algorithm sifts through the data to determine the essential elements of a picture that correspond to *cat* or *not a cat*. When confronted with a new picture, the inference algorithm can then come up with the likelihood that there is a cat in the photo, without human intervention.

Based on advancements in machine learning, artificial intelligence is making great leaps forward for consumers and businesses, in applications such as natural speech recognition and autonomous driving. Within the past few years, machines have beaten human experts at games including *Jeopardy!*, Go, and poker.

Today machine learning and AI are predominant in the domain of hyperscale, not HPC, though there are many similarities. The leading researchers in AI are the dominant hyperscale companies, with one notable addition: IBM, which has invested heavily in its Watson technology for “cognitive computing.” Several major supercomputing sites are working on AI, and Japan has announced two publicly funded supercomputing projects with AI focuses.

AI has the potential to touch almost any industry, and there are some it may revolutionize. Some of the possibilities are:

- ***Medicine:*** This is the most “marketable” of AI advancements, as there is a popular demand for technology that can improve and extend people’s lives. AI can be used to look across wide databases of symptoms and patient histories to arrive at more accurate diagnoses, especially for rare conditions, and to design personalized courses of treatment. AI can also monitor symptoms to alert medical personnel to important changes in patients’ conditions.
- ***Finance:*** Although less popular than medicine as an avenue for AI, finance is a hotbed of machine learning, as financial services organizations have troves of data that can be used to optimize pricing and investments. Any financial mechanism—whether it is a credit card, a mortgage, or an insurance policy—has a price, often in the form of an interest rate. Rather than dividing customers broadly into

risk categories (known as "tranches"), financial institutions can use machine learning to analyze the individual risk of any consumer, business, or investment, as it changes over time in response to changing inputs. AI can also be used to optimize fraud detection (in credit card transactions, insurance claims, etc.) and to anticipate future changes in market conditions.

- ***Retail:*** Recommendation engines are already important tools in retail. Most of us have seen messages on our screens with suggestions, "Since you recently bought (or downloaded, or browsed for) *X*, you might also be interested in *Y*." AI can do this more intelligently, across multiple platforms, with a wider array of data, fast enough to include recommendations and special offers at the moment of sale. AI can also be used to analyze and predict sales trends, enabling better inventory management. In the case of food, this can also reduce spoilage.
- ***Defense:*** Many of the advancements made in AI in research and industry have applications in defense. Most notably, AI can directly improve the analytics of vast amount of intelligence data, which can better enable insights both offensively (where and when to go after a target) and defensively (detecting terrorist activity and predicting attacks). Autonomous drones, robots, and vehicles have obvious strategic military benefits. And AI capabilities could extend to both cybersecurity and cyberattack strategies, whether planning or preventing assaults based on hacking.

Quantum Computing

Quantum computing represents a potential discontinuous advancement in the way supercomputers could be built. In the current model, all computing is based on "bits"—switches that can be off or on, which translate to the 0s and 1s that are the foundation of computation and logic. Eight bits form a byte, which gathered together form the Gigabytes, Terabytes, Petabytes, and Exabytes of information coursing through supercomputers.

Quantum computers are built on subatomic quantum particles, which have the unusual and important characteristic of "superposition"—they can be in more than one state at the same time. These quantum bits, or "qubits" (pronounced "cubits") can be equal to 0 or 1, or both 0 and 1, or any probabilistic combination of 0 and 1. A quantum computer can therefore explore a vastly greater solution space, based on fewer inputs.

The potential of quantum computing has been touted for many years, including recognition at the national level. Beginning in 2002, the High Productivity Computing Systems (HPCS) initiative under the U.S. Defense Advanced Research Projects Agency (DARPA) recognized the need for advancements in supercomputing that would "fill the high-end computing technology and capability gap" until the future promise of quantum computing.[xviii]

There are formidable challenges to building a quantum computer, well beyond mere expertise in quantum dynamics. The quantum particles need to be isolated, held in place, controlled, and measured, free from cosmic interference. To achieve this in a research setting is daunting; to do it commercially is harder still.

In 2007, D-Wave announced the first commercial quantum computer. It is uses "quantum annealing" techniques, which some have criticized as not fulfilling the full promise of a supercomputer based on "quantum entanglement." In short, the quantum annealing technology is most useful for exploring independent paths in a solution space, as each qubit seeks a low-energy point. (Imagine dropping marbles over an uneven surface. Each marble will roll to a local low point. The more marbles you drop, the greater the chances that at least one of them will find a path that rolls to the lowest point on the entire surface. This point is the "optimal solution" for the space.) D-Wave has sold a few systems, including to Google and NASA Ames Research Center, who are collaborating on the technology. The current D-Wave systems scale up to 2,000 qubits.

In March 2017 (less than two weeks before this testimony date), IBM announced the commercial availability of a "universal quantum computer," which utilizes "pairwise quantum entanglement," meaning any two qubits can be linked to each other's states. According to IBM, a system of *n* qubits is therefore capable of fully exploring a solution space with 2^n possibilities. Its current system of 5 qubits can therefore be programmed to look at 32 (2^5) possibilities in a space. If that seems small, consider that IBM also says it will have systems with up to 50 qubits "within a few years," which would be on pace to exceed current supercomputing capabilities for some categories of problems, *if* the technology works and it can be programmed.

Even in the most aggressive case, quantum computing will not displace conventional computing in the current planning horizon. At best it will provide a way to solve problem types that are not well-suited to current computing architectures. The first areas for quantum computing will be applications that take relatively few inputs, explore a large possible set of solutions, and provide few outputs. Playing chess is a metaphor for this type of application, but there are potential real-world scientific applications in materials science, biotechnology, and possibly also encryption (or decryption, as quantum computing might excel at factoring large numbers).

HPC INVESTMENT IN THE U.S., CHINA, AND REST OF WORLD

Throughout most of the history of supercomputing, the U.S. has been the dominant leader. Most of the vendors of key technologies have been U.S. companies, and U.S. national labs have been at the forefront of supercomputing research. Throughout industry and academia, the U.S. dominates in HPC usage.

However, access to HPC technologies is no longer rarified. Anyone with the budget (including capital budget, facilities, and human expertise) can build a supercomputer, from component technologies that are readily available. In the past, the U.S. was able to limit access to supercomputers through export controls. Today, individual technologies can be controlled, but there are alternatives available.

The Chinese program to build its national supercomputing efforts comes at a time when the model of how to build supercomputers is changing. With a paradigm shift, it is possible for one region to leapfrog ahead of another, even when it has been behind. As a metaphor, consider the revolution with mobile communications. A nation with limited telephone landline infrastructure could suddenly bound ahead in mobile with a well-timed investment. Such is the potential with supercomputing today. If Exascale systems look different from current systems, particularly in how they are programmed, then it is possible to essentially “come from behind” with timely investment. This section looks at the current supercomputing policy, strategy, and investment in the U.S., China, and the rest of the world.

Supercomputing in the United States

At a national level, current investment in new supercomputing technologies flows predominantly through DOE, including the DOE Office of Science and the NNSA. DOE national labs have been a focal point for supercomputing for decades, and that focus continues through the U.S. Exascale Computing Project.

A current collaboration known as CORAL (Collaboration of Oak Ridge, Argonne, and Livermore) is pursuing three “pre-Exascale” systems, using two different approaches. The supercomputers at Oak Ridge National Laboratory (ORNL) and Lawrence Livermore National Laboratory (LLNL) are based on IBM POWER technology and accelerated with NVIDIA GPUs[xix], while the deployment at Argonne National Laboratory (ANL) uses Intel technology, with Cray as a system integration partner.[xx] All three systems are planned to deliver over 100 Petaflops of performance, with target power budgets of 10 Megawatts.

These deployments highlight the competitive rift that has formed between the U.S. supercomputing vendors, which are now in two predominant camps. Intel is at the nexus of one camp, with system integration partners that will build scalable systems on Intel technologies, including its processing elements (both CPUs and FPGAs), interconnects (OmniPath and Ethernet), and accompanying software elements. The other camp is centered on IBM, which has its own systems and processors, and partners with other natural competitors to Intel in areas like accelerators (NVIDIA GPUs) and networking (Mellanox InfiniBand). Most technology vendors will feel pulled toward one camp or the other.

Current Program Investments

In the pursuit of Exascale technologies, the DOE Office of Science has synthesized the Exascale Computing Project (ECP). (The ECP supersedes the previous Exascale Initiative, also under the Office of Science, which included “Design Forward” and “Fast Forward” grants for companies designing next-generation technologies that could enable Exascale computing.)[xxi] As noted above, the U.S. had planned to field its first two Exaflop systems in 2022 to 2023, based on the CORAL “pre-Exascale” architectures, approximately three years behind China. A newly ratified plan would deliver an Exaflop system in 2021, based on a “novel architecture,” presumably different from the two CORAL architectures already planned. (ARM architecture is a strong possibility, which would be similar to Japan's plan, in roughly the same timeframe.)

The ECP has stressed Exa*scale* over Exa*flops*, with focus areas in hardware technology, software technology, application development, and Exascale systems[xxii]. For the CORAL systems and their follow-ons, the rationale has been that if the U.S. is not going to reach an Exaflop first, then it will at least do it better, with more efficient, general-purpose systems serving a wider range of scientific applications. While this argument is not without merit, it has also been subject to debate, ceding a multi-year head start to other countries, which would not hold still in the interim.

The new novel-architecture plan brings the U.S. timeline closer to China and Japan, but questions arise as to whether any Exa*scale* considerations are being sacrificed with the new machine. A new architecture will have less software readiness than an established one, and it is being introduced a year sooner. Our analysis is that the added deliverable will reduce pressure for the U.S. to deliver an Exa*flop* sooner, while still allowing the CORAL successors to move forward.

Additional national funding for HPC in the U.S. flows through the Department of Defense (DoD), the National Science Foundation (NSF), the National Oceanographic and Atmospheric Administration (NOAA) under the U.S. Dept. of Commerce, NASA, and the National Institute of Health (NIH).

The DoD HPC efforts, through the DoD HPC Modernization Program, are consolidated primarily into four major DoD Supercomputing Resource Centers (DSRCs): U.S. Army Research Laboratory, U.S. Naval Research Laboratory, Wright Patterson Air Force Base, and the Engineer Research and Development Center (ERDC) of the U.S. Army Corps of Engineers. Previously two additional DSRCs received funding—the Maui HPC Center (MHPCC) in Hawaii and the Arctic Region Supercomputing Center (ARSC) in Alaska—but these were dissolved in 2015.

NSF and NIH grants fund HPC at academic sites. Some of the largest and most significant academic supercomputing labs in the U.S. are the National Center for Supercomputing Applications (NCSA) at the University of Illinois, the Texas Advanced Computing Center (TACC) at the University of Texas, and the Pittsburgh Supercomputing Center, which is a joint collaboration between Carnegie-Mellon University and the University of Pittsburgh.

Changes in funding can of course happen at any time, particularly with shifts in power between political parties in Congress or in the White House, or in response to changes in national or global economic conditions. That said, supercomputing programs have often enjoyed bipartisan support in the U.S., due to the multi-purpose links to scientific research, to industrial advancement, and to national security.

Foreign Vendors in the U.S.

The U.S. supercomputing efforts are not reliant to any significant extent on technology from vendors based outside the U.S. A complete, best-of-breed supercomputer can be built exclusively from technology from U.S.-based vendors, with multiple options. That said, the major vendors are all multi-national companies. Intel, IBM, HPE, and Dell EMC (among others) all have major operations in China. Processing elements are often fabricated in China.

At the system level, foreign-based companies like Lenovo and Huawei have trouble competing for U.S. government-funded supercomputing acquisitions. However, they compete on more even footing for commercial HPC business.

Supercomputing in China

As established above, China has undergone a surge in supercomputing investment in recent years, to the forefront of achievement worldwide. China currently hosts the two most powerful supercomputers in the world by LINPACK (TOP500) performance, and the Sunway TaihuLight is five times more powerful than the most powerful supercomputers in the U.S. Put another way, the top two supercomputers in the U.S. *combined* achieve only roughly the same performance as the *second-most* powerful system in China.

The Chinese market is notoriously difficult to penetrate—even to monitor it, let alone to sell into it. It is neither a free-press nor free-speech society, and we are often left to make inferences from what we can observe. It is possible that China has other classified supercomputers that are not known to the TOP500 list or to the world at large. (For that matter, this is possible in the U.S. or other countries too.) But in the case of the Chinese market, we believe it is more likely that the supercomputers we know about are also serving government interests. A supercomputer that is configured and advertised to serve weather and climate modeling, for example, might also serve nuclear simulations. Intersect360 Research assumes that these Chinese national supercomputing centers serve both defense and scientific research purposes.

Current Program Investments

The Chinese supercomputing programs are neither as well-communicated nor as well-known as their U.S. counterparts. The Chinese government usually works on 10-year cycles. We initially questioned whether the current Politburo would support supercomputing to the extent of the previous one, which funded the original Tienhe ("Milky Way") supercomputer and its follow-on, Tienhe-2, which was installed after the changeover in power. Our assessment is that the level of investment has been at least stable, and may be increasing. Under its current administration, China seems determined to be the world leader in supercomputing technologies.

As part of this strategy, China is intentionally decreasing its reliance on imported technologies. The Tienhe supercomputer used Intel processors, integrated by Inspur. (Within China this was viewed as a Chinese system; outside China it was viewed more as an Intel system integrated in China.) Recently the U.S. government blocked certain processing technologies from export to Chinese supercomputing labs, due to the revelation that they were indeed conducting nuclear simulations.[xxiii] These export restrictions included the Intel processors slated for Chinese supercomputing upgrades.

This action by the U.S. government may have had unintended consequences. The planned upgrades were certainly delayed, but in the interim, the Chinese government increased its focus on domestic microprocessor technologies, including the Sunway processor. It is difficult to say for certain whether the TaihuLight system would be more powerful or more efficient using Intel processor technology, but what is certain is that the Chinese initiatives can no longer be thwarted by U.S. export control.

The area in which China may lag the furthest behind other countries is networking. As noted above, the Sunway TaihuLight system incorporated networking chips purchased from Mellanox. Although the official statement is that the Chinese developed a unique network based on those chips, we assume that the technology is effectively InfiniBand, or something very much like it, and that the Chinese government could not have built an efficient, high-performance system interconnect without these chips, at least not in a similar timeframe.

The most important distinction of the Sunway TaihuLight may be that it was *built*, not *bought*. That is to say, the supercomputing centers themselves designed the components and integrated them. It is not clear to what extent the resulting systems might eventually be commercialized for sale inside or outside China. This presents an interesting ramification to consider: When private U.S. companies design supercomputing systems or technologies, they seek to sell them in as broad a market as possible, inside or outside the U.S. Many other countries thereby benefit from the investment. But China is a relatively closed economy, and if the Chinese government designs a supercomputer that is better than any other, even if only for selected purposes, then it is not certain that technology will ever be available to anyone else.

One final point about the Chinese supercomputing strategy is that over time is has leaned more toward centralization than decentralization. Rather than pursuing a strategy that would put HPC resources into individual companies and researchers, we find it likelier that the Chinese government would create programs of access to centralized resources.

Foreign Vendors in China

U.S.-based technology vendors perceive China as a market with tremendous growth potential, and many have invested in strong Chinese presences to capitalize on it. The competitive dynamic mirrors that in the U.S. Companies like IBM, HPE, and Dell EMC can compete for corporate business, but they have little access to government bids versus system vendors based in China. Inspur is the market share leader for HPC systems deployed in China.

Notable Supercomputing Strategies in Other Countries

The U.S. and China are not the only two superpowers in supercomputing. This section provides a brief analysis of some other considerations at national levels.

Japan

Not long ago, Japan (not China) was the dominant supercomputing force in Asia, and Japan is still among the world leaders today. From 2002 to 2004, Japan’s “Earth Simulator,” a multi-agency supercomputer built by NEC and designed specifically for global climate and ocean modeling, was recognized as the most powerful supercomputer in the world. More recently, the “K” supercomputer, built by Fujitsu at the RIKEN Advanced Institute for Computational Science, was the world’s fastest supercomputer in 2012, and the first supercomputer to top 10 Petaflops on LINPACK. The K computer is still the seventh-fastest in the world in the current ranking.

Today Japan is charting its path toward Exascale computing with a “Post-K” architecture from Fujitsu. Previous Fujitsu supercomputers have been based on SPARC processors, a variety of RISC 64-bit processors pioneered and promoted by Sun Microsystems. Moving forward, the Post-K systems will be based on ARM processors, which can be viewed as native Japanese technology since SoftBank’s acquisition of ARM Holdings in 2016. Fujitsu is also noteworthy in that it continues to pursue its own custom interconnect technology, called “Tofu,” rather than relying on Mellanox InfiniBand, Intel OmniPath, or another networking technology.

Japan at one point announced plans to deploy an Exascale system in 2019. Recent delays make 2020 or 2021 a more likely target. At this pace, Japan will likely be the second country with a supercomputing lab at an Exaflop, after China.

France

Among European countries, France is most notable for having a native supercomputing vendor, Bull (now a brand within Atos), that is capable of fielding an Exascale system in the next few years. The likeliest customer would be CEA, the French national atomic energy commission, though Bull could as easily sell its systems in other European countries. At one point, public statements implied that such a system might be deployed as early as 2020.

The Bull architecture leverages processing elements from Intel, but like Fujitsu in Japan with its Tofu interconnect (and Cray in the U.S., with its Aries interconnect), Bull is pursuing its own system networking technologies, called BXI (for Bull eXtreme Interconnect).

Russia

Russia is worthy of special inclusion specifically because it has a native vendor, RSC Technologies, that has stated it would be capable of fielding an Exascale system in a timeframe similar to other national initiatives, if it had a customer. The likeliest buyer would be Moscow State University, which currently hosts the most powerful supercomputer in Russia, but we know of no confirmed, funded plan or timeframe. The RSC architecture uses Intel processors and no custom interconnect. Within Russia, RSC competes with T-Platforms, which previously dominated the Russian supercomputing market, but which was set back greatly by a temporary U.S. ban on its use of American-influenced technologies.

CONCLUSIONS AND RECOMMENDATIONS

Supercomputing is vital not only to scientific research, which benefits the world, but also to the U.S. economy, in areas such as manufacturing, energy, pharmaceuticals, and finance, as well as to national security interests. For most of the history of the HPC market, the U.S. has not only been dominant in both the usage and the production of supercomputing technologies, but also it has had the ability to selectively limit the export of supercomputing capabilities to other countries if desired.

As supercomputing architectures are evolving and specializing, new paradigms need to be developed along with them. Exascale systems will require different programming, different management, and different stewardship than their predecessors. The hyperscale market will influence this path, as will the attendant rise of artificial intelligence. In short, the rules of the game are changing.

The U.S. should not underestimate the capability or potential of Chinese supercomputing initiatives. If the systems are built, many brilliant scientists will go to work to find innovative ways to use them for scientific research, which in turn begets advancement in other areas. Furthermore, the U.S. administration should see that attempts to limit Chinese development through export regulation have been counterproductive; the immediate result was to boost China's own domestic technologies.

While the U.S. still leads by far in the most straightforward market share metrics of production (vendors, supply-side) and consumption (buyers, demand-side), industry indicators show the U.S. is falling behind in the leading edge of advancement, and simultaneously losing the ability to rein in other countries via export control. Chinese leadership has apparently recognized the relationship between HPC and economic growth and has set forth on a program to drive the country into a leadership position. The best response to this new challenge is to continue if not increase national support for HPC at all levels.

The great strength of the U.S. is its economy and the strength of its private sector. National supercomputing efforts are essential to motivating investment at the high end. From that point, U.S. companies excel at seizing opportunities to drive markets forward.

Against these strengths, the top limitations to Exascale deployments are software and skills. If we do build a system, how will we use it? A key feature of the ECP is its emphasis on co-design, finding end-user stakeholders to collaborate on the design of next-generation supercomputing technologies, bolstered by government funding.

Beyond the continuance of ECP, we offer the following recommendations:

- *National initiatives in low-level software tools and programming models, together with stakeholders in industry and academia.* While individual applications must be tailored to specific architectures and may be of interest to a limited audience, there is some software functionality that would be of broader benefit. This includes programming models—the methodologies with which application engineers get their ideas to scale on large machines—and work on common tools and algorithms, such as math libraries, which benefit multiple applications. This type of work is not flashy—it is hard to get the public excited about linear algebra implementations—but the downstream benefit to multiple domains is a major payoff.
- *Government-funded partnerships between industry and academia.* The skills gap is a significant problem across the HPC industry. There is a scarcity of engineers, and much of the talent is more attracted to work in hyperscale industries. If more HPC skills were introduced in academic science and engineering programs, government programs could connect the students with organizations in need of

their skills. Coursework could be designed such that the student is working on actual models as part of their course of study, at no cost to the participating company or organization. By the time the students join the workforce, they have learned valuable skills, become knowledgeable in a potential hirer's products or process, and provided a valuable service, even if they choose to do something else. For participating organizations, they get access to HPC skills (albeit entry-level ones) with little to no cost or risk, and access to a pre-trained employee if they choose to hire.

- *Ongoing pursuit of next-generation technologies.* As noted throughout this statement, supercomputing is an industry of change. Beyond the leading vendors mentioned at various points, there are countless startups in pursuit of game-changing ideas, one of which might turn the industry on its head in five, ten, or 20 years.

Regardless of these recommendations, the HPC market will continue, powering new innovations and ideas around the world. Supercomputers today are close to a million times more powerful now than they were 20 years ago. In another 20 years, they could be a million times more powerful still. The leaders in supercomputing will be the ones that do not rest on their achievements, but rather continue to chase the next challenge over each new horizon.

APPENDIX A: QUESTIONS FROM USCC

Mr. Snell has submitted this statement in response to the following questions from USCC. Intersect360 Research Chief Research Officer Christopher G. Willard, Ph.D., contributed to this analysis. Data from Intersect360 Research surveys and forecasts has been included where relevant.

Question 1

Briefly describe the current status of high performance computing including both hardware and software and its applications. What is driving developments in these areas? How are artificial intelligence, next-generation semiconductors, deep learning, and big data playing a role in further advancements? How will advancements in high-performance computing affect a country's military capabilities and global competitiveness?

Question 2

Compare and contrast U.S. and Chinese technological capabilities and pace of innovation in high-performance computing. To what degree does Chinese high-performance computing demonstrate improvements over foreign systems and breakthroughs in technology? In what areas, is the United States still a technological leader?

Question 3

Briefly describe China's major industrial policies and plans supporting the development of its high-performance computing sector. How is the Chinese government implementing these plans and building its domestic capabilities in high-performance computing? What kinds of support (financial, regulatory, etc.) has the Chinese government provided to its domestic firms, research institutes, and universities? How much funding has the central and local governments allocated to support this sector? What is the role of technology transfer, licensing and other arrangements for sharing intellectual property, overseas investments (angel, greenfield, etc.), acquisitions (mergers and acquisitions or joint ventures), and recruitment of leading academics and overseas talent in enhancing China's advancements? Overall, how successful have those efforts been? What are the remaining challenges?

Question 4

Assess U.S. firms' operations in China and U.S. firms' business strategies to supply China's high-performance computing sector. What share of Chinese high-performance computing market do U.S. firms account for? How dependent are U.S. firms on the Chinese market? Which markets are the greatest opportunities for foreign market participation, and why? Which markets are the most restrictive, and why? Do foreign firms face any unfair or discretionary limitations (e.g., localization requirements, regulations, etc.) or technology transfer expectations? How are U.S. and other foreign firms coping with those restrictions, and what, if anything, should the U.S. government do about it?

Question 5

Assess the implications of China's high-performance computing development for the United States. How will these developments affect U.S. global competitiveness and technological edge? How will these developments affect U.S. military superiority and U.S. power projection capabilities? What restrictions (export controls, etc.) has the U.S. government placed on U.S. firms competing in China's market, and are these restrictions necessary? How effective have these restrictions been? How should the U.S. government balance its commercial and national security interests in high-performance computing?

Question 6

Assess how the United States can maintain its strategic advantage in high-performance computing going forward. How has the U.S. government supported the development of high performance computing in the United States? How could the U.S. government help the United States maintain its strategic advantage in high-performance computing and ensure high-paying jobs in these fields and research and development centers are located in the United States?

Question 7

The Commission is mandated to make policy recommendations to Congress based on its hearings and other research. Assess the implications of China's high-performance computing for United States. What are your specific recommendations for legislative and administrative action?

APPENDIX B: ABOUT INTERSECT360 RESEARCH AND ADDISON SNELL

About Intersect360 Research

Intersect360 Research was founded in January 2007 as a market intelligence, research, and consulting advisory practice focused on suppliers, users, and policy makers across the High Performance Computing ecosystem. The company operated as Tabor Research, a division of Tabor Communications, until it was purchased from the parent company by its two top executives, Addison Snell and Chris Willard, in 2009.

Intersect360 Research's deep knowledge of HPC, coupled with strong marketing and consulting expertise, results in actionable intelligence for the HPC industry—insights and advice that allow clients to make decisions that are measurably positive to their business. The company's end-user-focused research is inclusive from both a technology perspective and a usage standpoint, allowing Intersect360 Research to provide its clients with total market models that include both traditional and emerging HPC applications.

Intersect360 Research participates in the Advanced Computing Roundtable (previously the “HPC Advisory Committee”) of the U.S. Council on Competitiveness, and partnered with the Council on its report, “Solve. The Exascale Effect: The Benefits of Supercomputing Investment for U.S. Industry.” For “Solve,” Intersect360 Research conducted both quantitative and qualitative research of HPC-using companies to produce a comprehensive report linking federal supercomputing investment to industrial innovation. The “Solve” report can be downloaded from the Council web site at www.compete.org/reports/all/2695-solve.

In addition to its market advisory subscription services, Intersect360 Research offers an array of client-specific services, including custom surveys, white papers, custom analysis, and both marketing and general business consulting. More information on Intersect360 Research is available at www.intersect360.com.

About Addison Snell

Addison Snell is the CEO of Intersect360 Research and a veteran of the High Performance Computing industry. He launched the company in 2007 as Tabor Research, a division of Tabor Communications, and he brought the company independent in 2009 as Intersect360 Research together with his partner, Christopher Willard, Ph.D. Under his leadership, Intersect360 Research has become a premier source of market information, analysis, and consulting for the HPC and hyperscale industries worldwide. Mr. Snell was named one of 2010's “People to Watch” by HPCwire.

Prior to Intersect360 Research, Mr. Snell was an HPC industry analyst for IDC. He originally gained industry recognition as a marketing leader and spokesperson for SGI's supercomputing products and strategy.

Mr. Snell holds a master's degree from the Kellogg School of Management at Northwestern University and a bachelor's degree from the University of Pennsylvania.

ENDNOTES

[i] For further reading on this topic, see: Snell, Addison, “Beyond Beowulf: Clusters, Cores, and a New Era of TOP500,” published on TOP500.org, November 11, 2014, https://www.top500.org/news/beyond-beowulf-clusters-cores-and-a-new-era-of-top500/.

[ii] Intersect360 Research special study, “Processor Architectures in HPC,” 2016.

[iii] Because the *ps* at the end of the word *flops* stands for “per second,” some people also use “flops” as a singular noun—one flops, one Petaflops. Colloquially the singular is often shortened to “flop,” and we used this construct in this statement.

[iv] Intersect360 Research, *This Week in HPC* podcast, June 21, 2016, https://soundcloud.com/this-week-in-hpc/new-1-supercomputer-crushes-competition-and-china-takes-top500-by-storm.

[v] To view the current (and past) TOP500 rankings, along with analysis of the systems on the lists, see http://www.top500.org.

[vi] Specifically, the High Performance LINPACK (HPL) benchmark; for more information, see: https://www.top500.org/project/linpack/.

[vii] For this and other market revenue figures, we use a 2016 forecast based on 2015 actuals. Intersect360 Research will complete its analysis of 2016 actual revenue in roughly May 2016.

[viii] Intersect360 Research, “Worldwide High Performance Computing (HPC) 2015 Total Market Model and 2016–2020 Forecast,” September 2016.

[ix] Intersect360 Research forecast data, 2016.

[x] Intersect360 Research forecast data, 2016.

[xi] Intersect360 Research forecast data, 2016.

[xii] Intersect360 Research, *This Week in HPC* podcast, June 28, 2016, https://soundcloud.com/this-week-in-hpc/knights-landing-and-pascal-gpu-face-off-at-isc-and-fujitsu-surprises-with-arm

[xiii] Intersect360 Research, “The Hyperscale Market: Definitions, Scope, and Market Dynamics,” April 2016.

[xiv] Intersect360 Research special study, “The Big Data Opportunity for HPC,” 2012 and 2013.

[xv] U.S. Council on Competitiveness and Intersect360 Research, “Solve. The Exascale Effect: The Benefits of Supercomputing Investment for U.S. Industry,” October 2014, http://www.compete.org/reports/all/2695-solve.

[xvi] Intersect360 Research, This Week in HPC podcast, December 13, 2016, https://soundcloud.com/this-week-in-hpc/episode155-hpe-puts-the-machine-in-motion-us-embarks-on-faster-path-to-exascale.

[xvii] Feldman, Michael, TOP500.org, “First US Exascale Supercomputer Now On Track for 2021,” December 10, 2016, https://www.top500.org/news/first-us-exascale-supercomputer-now-on-track-for-2021/.

[xviii] Graybill, Robert, DARPA/IPTO, “High Productivity Computing Systems,” March 13, 2003, https://science.energy.gov/~/media/ascr/ascac/pdf/meetings/mar03/Graybill.pdf.

[xix] Intersect360 Research, *This Week in HPC* podcast, November 16, 2014, https://www.top500.org/news/ibm-and-nvidia-pick-up-two-pieces-of-coral-and-a-look-ahead-to-sc14/.

[xx] Intersect360 Research, *This Week in HPC* podcast, April 13, 2015, https://www.top500.org/news/intel-gets-final-piece-of-coral-altera-deal-in-doubt/.

[xxi] Exascale Initiative web site, http://www.exascaleinitiative.org.

[xxii] Exascale Computing Project web site, https://exascaleproject.org/exascale-computing-project/.

[xxiii] Intersect360 Research, *This Week in HPC* podcast, April 20, 2015, https://www.top500.org/news/us-drops-bomb-on-chinese-supercomputing-export-restrictions-threaten-tianhe-2-expansion/.