National Science Foundation
4201 Wilson Boulevard
Arlington, Virginia 22230

NSF Report on Support for Cloud Computing

In response to America Competes Reauthorization Act of 2010, Section 524

Preface

NSF is pleased to submit this report to Congress on the outcomes of National Science Foundation investments in cloud computing research, recommendations for research focus and program improvements, and other related recommendations, as required by section 524.c of the America Competes Reauthorization Act of 2010.

NSF appreciates the considerable discretion given by 524.b.2, allowing NSF to define programs of research and NSF's Program Officers to seek and manage research projects in Cloud Computing according to their best scientific judgments.

Section 524 of the America Competes Reauthorization Act of 2010 is as follows:

SEC. 524. CLOUD COMPUTING RESEARCH ENHANCEMENT.

(a) Research Focus Area- The Director may support a national research agenda in key areas affected by the increased use of public and private cloud computing, including--

(1) new approaches, techniques, technologies, and tools for--

(A) optimizing the effectiveness and efficiency of cloud computing environments; and

(B) mitigating security, identity, privacy, reliability, and manageability risks in cloud-based environments, including as they differ from traditional data centers;

(2) new algorithms and technologies to define, assess, and establish large-scale, trustworthy, cloud-based infrastructures;

(3) models and advanced technologies to measure, assess, report, and understand the performance, reliability, energy consumption, and other characteristics of complex cloud environments; and

(4) advanced security technologies to protect sensitive or proprietary information in global-scale cloud environments.

(b) Establishment-

(1) IN GENERAL- Not later than 60 days after the date of enactment of this Act, the Director shall initiate a review and assessment of cloud computing research opportunities and challenges, including research areas listed in subsection (a), as well as related issues such as--

(A) the management and assurance of data that are the subject of Federal laws and regulations in cloud computing environments, which laws and regulations exist on the date of enactment of this Act;

(B) misappropriation of cloud services, piracy through cloud technologies, and other threats to the integrity of cloud services;

(C) areas of advanced technology needed to enable trusted communications, processing, and storage; and

(D) other areas of focus determined appropriate by the Director.

(2) UNSOLICITED PROPOSALS- The Director may accept unsolicited proposals that review and assess the issues described in paragraph (1). The proposals may be judged according to existing criteria of the National Science Foundation.

(c) Report- The Director shall provide an annual report for not less than 5 consecutive years to Congress on the outcomes of National Science Foundation investments in cloud computing research, recommendations for research focus and program improvements, or other related recommendations. The reports, including any interim findings or recommendations, shall be made publicly available on the website of the National Science Foundation.

(d) NIST Support- The Director of the National Institute of Standards and Technology shall--

(1) collaborate with industry in the development of standards supporting trusted cloud computing infrastructures, metrics, interoperability, and assurance; and

(2) support standards development with the intent of supporting common goals.

---

1. Introduction: The Importance of Cloud Computing

The National Institute of Standards and Technology (NIST) has identified Cloud Computing as a vital area of national importance that requires further research and development; see http://www.nist.gov/itl/cloud/index.cfm.

"The Cloud Computing model offers the promise of massive cost savings combined with increased IT agility. It is considered critical that government and industry begin adoption of this technology in response to difficult economic constraints. However, cloud computing technology challenges many traditional approaches to datacenter and enterprise application design and management. Cloud computing is currently being used; however, security, interoperability, and portability are cited as major barriers to broader adoption."

NIST has defined Cloud Computing as follows: (see http://csrc.nist.gov/publications/drafts/800-145/Draft-SP-800-145_cloud-definition.pdf)

> "Cloud Computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models. The five essential characteristics are on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service. The three service models are Cloud Software as a Service (SaaS), Cloud Platform as a Service (PaaS), and Cloud Infrastructure as a Service (IaaS). The four deployment models are Private Cloud, Community Cloud, Public Cloud, and Hybrid Cloud." The reader is referred to the full text of the NIST definition for details of these elements.

The increasingly popular cloud model provides as-needed access to computing with the appearance of unlimited resources. Users are given access to a variety of data and software utilities to manage their work. Billing for services is based on usage; users rent virtual resources and pay for only what they use. Underlying these services are data centers that provide "virtual machines" (VMs). Virtual machines make it easy to host computation and applications for large numbers of distributed users by giving each the illusion of a dedicated computer system. It is anticipated that cloud platforms and services will increasingly play a critical role in academic, government and industry sectors, and will have widespread societal impact. Accordingly, NSF is committed to providing the science and engineering communities with the opportunity to conduct research and education activities in cloud and data-intensive computing and their applications.

2. CISE Cloud Computing Awards 2009- 2011

Research in the new area of Cloud Computing extends the research conducted in many older and more established areas. The areas are not mutually exclusive, and many projects contribute to more than one. We emphasize the following areas in this report:

1. Computer Systems
2. Computer Networks
3. Security and Privacy
4. Algorithms and data management
5. Applications and software engineering
6. Computer science education

As required, the Computer and Information Science and Engineering (CISE) Directorate of NSF has, as an ongoing effort, reviewed and assessed cloud computing research opportunities and challenges. CISE is increasingly funding awards in these six areas that are specifically targeted to or have significant impact on cloud computing, including the key topics listed in section 524. These cloud computing projects include 76 active awards managed by the Computer and Network Systems {CNS) Division, 40 active awards managed by the Computing and Communications Foundations (CCF) Division, and 9 active awards managed by the Information and Intelligent Systems {IIS) Division. Titles, amounts, and abstracts of these 125 awards will be made available upon request. These awards represent approximately 2.7% of the new awards made by CISE in FY09-FY11: 4.1% of new awards in CNS, 3.1% in CCF, and 0.6% in IIS.

In this section we describe some of the cloud computing projects within these important areas that are currently supported in the CISE portfolio.

2.1Cloud Computing Awards in Computer Systems

These awards explo re the following research directions:

1. Cloud Architectures:

   One project is exploring different design paradigms for data center networks. In particular, it pushes functionality from the network switches into the servers, which generally have more standardized hardware and software platforms, and uses multiple layers of low-cost, commercial-off-the-shelf (COTS) mini-switches to connect the servers. The project leverages the large number of servers in a data center, as to create a dense network of connections between them without using high-end switches. This dense network provides the foundation for enhancing the scalability, inter-server capacity, and fault tolerance of the data center network.

Another project is exploring new memory architectures for data centers. Instead of relying solely on the technology of dynamic random access memory (DRAM) commonly used today, it uses multiple technologies, such as phase-change memory (PRAM) and flash memory, to construct a high-capacity, energy- efficient memory system for computer servers in the clouds. This research makes fundamental contributions to computer science in the development of integrated memory architectures that can seamlessly manage hybrid memory resources, as well as the creation of novel algorithms and policies for energy, performance and endurance management.

Other projects focus on the new generations of cloud servers that will be implemented using multi-core chips, i.e., chips that have a large number of core processors and caches. These projects are currently designing and analyzing different methods and algorithms that can be used to speed up the execution of cloud server software on multi-core chips.

2. Green Clouds:

According to the U.S. Environmental Protection Agency, data centers in the United States incur an annual energy cost of $4.5 billion, which is comparable to the total consumption of 5.8 million average US households. Moreover, in the absence of intervention, the total energy consumption of data centers is expected to double every five years. Fortunately, up to 80% of this projected energy expenditure may be avoidable, which adds up to a prospective reduction in nationwide carbon dioxide emissions of 47 million metric tons (MMT) per year.

To deal with these problems, some projects are developing techniques for increasing the energy-efficiency of modern data centers while maintaining computational performance. Both computing and cooling energy are considered in these techniques. Many of these techniques are based on the following two observations. First, the rapid growth in hardware parallelism, such as multi-core chips, leaves more unused "residual resources" to be exploited. Second, one can use these residual resources to run a secondary computation in the background while a primary foreground computation is running. This background process can be added with a relatively low increment in power usage.

Additionally, many energy-efficiency techniques try to manage idle time on servers. Data center servers are only busy 10-30% of the time on average, but they are often left on while idle. During this time, servers can use 60% or more of their peak power consumption. Therefore, some projects are developing simple "load-oblivious" policies that can automatically scale data center resources to reduce power consumption. The developed policies change the number of servers that are switched on as a function of the changing workload. Servers are shut down, put into sleep states, or run at different frequencies, without knowing in advance how much work needs to be done.

Other researchers are investigating how to optimize the operation of data centers in the face of carbon emissions trading frameworks (e.g., cap-and-trade/Kyoto-style, cap-and-pay frameworks, and unregulated) to help businesses optimize energy usage to save money or achieve carbon neutrality.

Another project is investigating how to design servers in a data center to run on variable power from renewable sources, such as solar and wind power. This project is investigating new techniques to manage and adapt the energy and power footprint of a server to fluctuations in supply. These techniques require a server to be "duty-cycled," i.e., turned on and off in a carefully controlled fashion, so that its power and energy footprint can be gracefully adapted to a varying supply curve.

2.2 Cloud Computing Awards in Computer Networks:

These awards explore the following research directions:

1. Network Support for Clouds:

Modern computer networks need to provide new communication primitives (basic operations such as naming, routing, switching, congestion control, and content distribution) that are more suitable for clouds. For example, a project is exploring the possibility of basing network operations (such as naming and routing) on service-level objects (such as applications or databases) instead of basing them on computer hosts.

Traditionally, the network of a datacenter is configured to use the same routing tree to choose the route each message will flow through, regardless of the application using and sending the message. New designs are being explored that allows each application in the datacenter network to create its own custom routing tree.

Another project is developing new "optimal swarming algorithms" for distributing content over multiple multicast trees, which delivers messages from the clouds to large client populations. These algorithms are able to communicate while achieving several desirable performance objectives.

Other researchers investigate how to employ "precision time protocols" to achieve time-predictable delivery of content from clouds to clients.

2. Marketplace for Clouds:

Researchers are exploring the merits of adopting "co-location games" as a framework for how clouds can be deployed and used in an economically sound manner. These games consider how

rational, self-interested parties interact to secure their share of cloud resources to support their applications. They model and analyze the dynamics that result when these parties negotiate to minimize their individual costs.

Another project investigates the challenges associated with migrating enterprise applications to clouds. These challenges include: (i) models to predict the impact of this migration on the performance of applications, (ii) methodologies to ensure that polices for network access to the applications on clouds are configured correctly, and (iii) methodologies to ensure that the applications, when executed on clouds, provide adequate quality of service guarantees.

Another project is working on evolving the client machines from traditional desktops that have dedicated hardware and software configurations into "virtual desktop clouds" that are accessible via lightweight client machines.

3. Resilient Networks for Clouds:

Researchers are working on quantifying the frequencies, durations, causes, and impacts of faults that occur within datacenter networks, as well as other network classes. The objective of this work is to generate a model of network faults that can be used to evaluate how suitable or effective different applications and protocols are for various datacenter network architectures.

With the move towards nano-scale fabrication, computer chips are increasingly becoming vulnerable to "soft errors" caused by external noise, and are increasingly likely to fail early due to fatigue. This trend has inspired a project to investigate computational "resilience methods" that scale gracefully in the face of increasing hardware failures. These methods will use novel partitioned redundancy strategies that achieve resiliency to network failures, and will function at different levels across hardware and software layers.

2.3 Cloud Computing Awards in Security and Privacy:

These awards address three broad thrusts:

1. Trustworthiness of Cloud Providers:

By utilizing cloud computing, organizations rely on third-party cloud providers to preserve their users' security and privacy. The effectiveness of cloud computing services is limited by the amount oftrust that users have in providers. As a result, much research is being conducted on how providers can provide more trustworthy services and how to prove their trustworthiness.

Cloud users can protect their data from being read or released by a cloud provider by encrypting it. Until recently, however, all known encryption methods would have resulted in data that

could not be manipulated at all by the cloud provider; the cloud would have been limited to providing nothing more than storage. This would have severely hampered the usefulness of the cloud. However, a recent NSF-funded breakthrough on {(homomorphic encryption" showed that it was theoretically possible for data to be encrypted but still allow a service provider to perform useful, trusted operations on it. These first results were impractical, but research is ongoing to make these general techniques practical. Other researchers are developing methods for specific services, such as database operations, to be provided securely and efficiently.

Other researchers are conducting work on {(functional encryption" to solve another practical cloud computing issue. Consider a company that uses encryption to store its sensitive information in the cloud. If each employee uses an individual encryption key, then when an employee dies or leave the company, the corporation risks losing all data encrypted by that individual. If, however, a company shares encryption keys amongst employees, then all such employees will have complete access to the data, regardless of later employment changes. Functional encryption gives another alternative, which is more appropriate for cloud computing. Data can be encrypted so that only those users currently holding appropriate credentials can access the data. For example, current members of an audit department can be granted full access, while other workers are restricted to view only data items relevant to their job functions.

Other projects are developing techniques for cloud computing providers to prove that they are appropriately protecting their users' information and enable users to audit providers for compliance.

2. Protecting cloud providers from threats:

Cloud providers face significant security and privacy threats, and research is being conducted to identify and protect against these threats.

By providing services to many organizations, a service provider becomes a high-value target of attacks. In addition to threats faced by traditional, independent computing departments, cloud computing providers can expect to face attacks that were previously economically infeasible to perform against traditional organizations because they are extremely difficult to mount. Given the large number of users in a cloud, such attacks become profitable when directed against a cloud provider. Cloud providers also have to protect their users from each other: if one client is a malware victim, other clients of the same provider should not be endangered. Also, an attacker should not be able to buy services from the same provider as its intended victim and then use those services to defeat the provider's defenses.

Various projects have explored architectures for cloud providers that systematically defend the provider from both external attacks and attacks mounted by their clients. These architectures are resilient to various attack mechanisms that affect traditional computer systems. They also

provide monitoring mechanisms that can allow a provider to detect and mitigate threats, and to recover from malicious activity.

In addition, NSF-funded researchers have discovered that under certain circumstances, some aspects of users' private information can be discovered by closely monitoring the activity of a cloud provider, and are developing mechanisms for defeating such attacks.

3. Leveraging the cloud to provide trustworthy applications:

Cloud computing has the potential to enable novel applications, and research on maximizing this potential in a trustworthy fashion is ongoing. Cloud computing not only offers cost savings, but also allows for new classes of applications that take advantage of increased information mobility and collaboration opportunities. Researchers funded by the Trustworthy Computing Program are attempting to realize these benefits.

For example, because health care records are held by many different organizations, important medical information is often not obtained in a timely fashion-a lapse that can result in injury or death. Cloud computing can help centralize health care records, provided that appropriate privacy and security mechanisms are in place to protect patients, doctors, and insurance providers. The Trustworthy Computing Program is funding work to develop these mechanisms.

Companies are ever-increasingly collaborating with each other, and cloud computing can serve as an enabling agent for the necessary information sharing. However, every company has its own internal security and privacy policies, and different companies might have different regulatory restrictions on data usage. Research is under way to allow cloud computing providers to understand and enforce these differing policies, so that each company can benefit from collaboration without the risk of violating its own regulations.

2.4 Cloud Computing Awards in Algorithms and Data Management:

These awards explore the following research directions:

1. Fundamental Work in Algorithms:

Cryptography, which was described in the section on Security and Privacy, is a major component of algorithm research in cloud computing. Other current fundamental algorithms research includes work in scheduling, resource management, coding theory, and formal methods.

Scheduling is a difficult and fundamental problem that pervades all of computer science, and cloud computing is no exception. New challenges arise from scheduling multiple instances of

multiple types of resources given multiple competing objectives. These objectives may include energy consumption as well as the service level agreements of multi-resident cloud users. For some workloads, storage requirements dominate, while for others, computational requirements dominate, and for still others, communication requirements dominate. A number of projects work on methods to resolve scheduling challenges; for some, it is the main theme of the research.

Virtualization, one of the cornerstones of cloud computing, maps physical resources into logical resources so that they can be shared in dynamic environments. The mapping process has an online component, because the system must respond to requests for resources as they arrive. In traditional online problems, resources are thought of as irrevocably assigned. However, in virtualized environments, previously allocated resources must be reassigned or reallocated as required. One project formalizes how computational resources can be reallocated efficiently given some cost for the reallocation. The project investigates the algorithmic complexity of reallocation problems that arises in large-scale systems design.

Understanding how to represent, store, and process information in distributed storage systems is a fundamental challenge in cloud computing. One project is designing novel distributed storage codes that use "network coding theory" to address modern storage challenges. This project seeks optimal strategies for storing large amounts of data in a distributed fashion, and to understand the dynamics of coded storage systems; i.e., how to maintain time-varying coded information representations over networks.

The correctness of different components of cloud computer systems can be verified using formal proof-theoretic frameworks. One project focuses on frameworks for three critical components in a typical cloud computing system: the partition management protocol (which provides the computing elasticity of the system), the storage subsystem, and the atomicity-guaranteeing protocol (which preserves integrity of data and computations).

2. Handling Massive Data:

Since one of the primary goals of the cloud is to host data-intensive applications, large-scale data management is a crucial component of cloud computing. Several projects explore the challenges in scaling up and scaling out database operations to the extreme data sizes and levels of distribution that will be present in the cloud environment. Some of these projects consider security issues raised by multiple cloud users sharing large scale data resources; others make energy consumption a primary area of concern.

Another project is developing theoretical foundations, as well as practical control algorithms, to enable scalable design and efficient management of future extreme-scale data-intensive computing. The researchers will identify fundamental design principles needed to achieve scalability for large scale network infrastructure and software systems. They will also develop

distributed control strategies on operator placement, data storage, load shedding, and resource allocation to allow for efficient in-network information processing. These advances require a collaborative effort spanning multiple disciplines including performance modeling, networking, queueing theory, and optimization.

In addition, one project is investigating how critical data-driven computer vision tasks, such as nearest neighbor searching and clustering in high-dimensional spaces, can be designed for cloud computing systems.

2.5 Cloud Computing Awards in Applications and Software Engineering:

These awards explore the following research directions:

1. Developing Applications for Cloud Environments:

A number of projects focus on developing particular applications or support for particular application domains in cloud computing environments. These diverse application domains include hydrology/water resource management, natural language processing on web-scale data sets, geographic information systems, biological applications such as protein folding and evolutionary biology, in-home control applications, and even theoretical physics such as computational string theory.

One project developed technology for ubiquitous event reporting and data gathering on the 2010 oil spill in the Gulf of Mexico and its ecological impacts. This project exploited the availability of smart phones (which have sophisticated sensor packages, high-level programming APIs, and multiple network connectivity options) and cloud computing infrastructures to enable collecting and aggregating data from mobile applications. The goal of the project was to develop a scientific basis for managing quality-of-service, user coordination, sensor data dissemination, and validation issues that arise in mobile disaster monitoring applications.

2. Software Engineering for the Cloud Environment

Several projects focus on improving the programmability of parallel and distributed systems. Some are developing better methods of detecting and remedying errors. Other projects consider programming and algorithm design models for cloud applications and for energy-efficiency in large-scale systems. Another project develops algorithms for ensuring that cloud computing systems provide (soft or hard) real-time guarantees.

One project is designing, implementing, and evaluating extensions to the services offered by cloud purveyors so that cloud services can be used by a broader developer base. For example, these improvements will enable scientists and students, who often require support for general

compute-intensive applications, to better work on the clouds. This project is planning to develop the fundamental technologies necessary for support of high-performance, compute intensive applications within the clouds. These technologies include efficient support of shared- memory inter-process communication, high-performance file system support, and system-wide performance monitoring and analysis tools.

There are two well-known computation models currently being used to program applications over the clouds. In the "in-memory model," intermediate states of the computation are held in the aggregate memory of the many machines in the cloud and shared and exchanged between them. In the "map-reduce model," the computed data simply flows between the different machines in the cloud and there are no stored intermediate states of computation. Each of these two models has advantages and disadvantages. Researchers are investigating how to combine these into a single model that benefits from the advantages of both.

Another project leverages a "group compositional approach" to tackle the problem of designing and programming large applications over the clouds. In this approach, the researchers first develop novel protocols for smaller groups of nodes which offer strong properties with minimal overhead. Second, they propose a coordination service and a suite of management algorithms to adaptively organize these smaller groups, composing them together into a large application. These management algorithms will address the problems of dynamic load balancing, topological control, and security.

Additionally, another project is investigating cloud computing systems serving multiple users with differing incentives. In particular, this project is developing ways to ensure the reliability of these systems in face of unresponsiveness or selfish behavior by the underlying competing users.

2.6 Cloud Computing Awards in Computer Science Education:

1. Parallel and Distributed Computing Curriculum

Parallel and Distributed Computing (PDC) now permeates most computing activities; this is especially true in the cloud computing environment. Certainly, it is no longer sufficient for even basic programmers to design only conventional sequential programs. This necessitates developers of all levels to have a broad-based skill set in parallel and distributed computing, which strongly impacts Computer Science (CS) and Computer Engineering (CE) programs as well as related computational disciplines. One project is currently developing a set of core topics in parallel and distributed computing for undergraduate CS and CE curricula.

2. Undergraduate Research in Cloud Computing

Currently, one Research Experience for Undergraduate (REU) site is focused on "Computer Systems Research in High Performance Cloud Computing Environments." Undergraduate researchers are immersed in departmental research areas of expertise, which includes computer architecture, energy-aware computing, virtualization, security, and cloud and grid computing.

2.7 Some Early Successes from the Current Award Portfolio

The currently funded portfolio of awards has already produced several early successes.

1. World record speed for data sorting

   One project set multiple world records in data sorting. On a cluster of only 50 servers, the project's team was able to break a 2-year old record for data sorting set by Yahoo Corporation, which ran on 3452 servers. Not only did the team manage the fastest sort time---lOOTB in 106 minutes---it also set the world record for most energy-efficient sorting. Data sorting is the foundation supporting a number of important high performance and data center activities, including web index generation, large-scale data analysis, and bioinformatics simulations. More detailed information on these sorting records is available at www.sortbenchmark.org.

2. Challenging conventional knowledge on achievable performance

   Another project has succeeded in challenging conventional wisdom regarding consistency and security of data storage in cloud computing.

   As storage systems have grown to Internet scales, they have often sacrificed consistency for performance due to the perceived costs and lack of scalability of consistency semantics. However, stronger consistency guarantees make correct programming for the system much easier. To resolve this issue, the project has developed several key-value storage systems that challenge the idea that a strict tradeoff must be made between consistency and performance. These include a system that is able to achieve higher throughput for common cloud workloads despite having the strongest notion of consistency (i.e., linearizability), as well as a system which can execute operations in the local datacenter with high availability and low latency. Such a system enables clients to obtain a consistent view of multiple keys through read transactions, and ensures a strong form of consistency when replicating data between datacenters in a scalable fashion.

   The history of data services is rife with unplanned data disclosures, malicious break-ins, and insider attacks. As described in detail in the Security section of this document, centralization of information makes cloud providers high value targets for attack. In addition to the trade-off between performance and consistency, the project also challenges the assumption that applications must sacrifice security (i.e., integrity) for privacy (i.e., confidentiality) in the cloud

environment. The project has been able to develop a series of systems that leverage less trusted cloud infrastructure. One system uses history to increase throughput by partially replicating work and basing correctness on consensus protocols (i.e., achieving security if at least some fraction of servers do not behave maliciously). A new algorithm can securely partition a large number of nodes into smaller, fault tolerant groups; even malicious attempts to thwart this partitioning mechanism are prevented. Finally, the project developed ways to build group collaboration applications even when the cloud is completely untrusted. Cloud servers only see clients' encrypted operations, and clients use cryptographic integrity checks to detect servers acting maliciously. If a malicious server adds, modifies, drops, or reorders updates, the clients can detect the server's misbehavior, switch to a new server, restore a consistent state, and continue in a correct state.

3. Fast and energy-efficient computer memory

Despite the advantages of cloud computing, energy consumption has become a critical challenge for data centers hosting the cloud. According to the EPA, data centers today consume up to 100 billion kilowatts of power. The bulk of power consumption has shifted from computer processors to computer memory with the development of efficient multi-cores and the increased software demands on memory. Unfortunately, today's memory technology (DRAM) is rapidly reaching its limit in power consumption and capacity for data-center sized applications.

As an alternative to traditional DRAM, which stores digital information with an electrical charge, one project explores a new technology called phase-change memory (PCM). PCM stores digital information by melting the crystalline structure of a material. This physical process requires no electrical power beyond the initial heating process to store information. However, PCM is relatively slow and wears out as the memory is used, much like the way automobile tires wear out and must be occasionally rotated for the longest tread life.

A project has shown that combining a small DRAM (for fast retrieval) with a larger and slower PCM (for storage of most information) results in a memory system that has drastically reduced power consumption while still being fast enough for most software program. This hybrid can also hugely increase the amount of information that can be stored. These innovations have led to an eight-fold reduction in power cost. Additionally, the techniques also greatly improve phase-change memory lifetime, which can now last long enough for several years of use in a data center.

The techniques and results from this project have influenced both academic and commercial approaches. Several companies, including Samsung, Rambus, IBM, and Micron, are investigating memory circuits and organizations that can prolong PCM lifetime by a factor of two or more and reduce memory energy consumption by more than 50%. The outcomes from the project may be incorporated in future commercial designs.

3. Initiatives on Cloud Computing 2009- 2011

We now describe several recent research efforts supported by CISE with a particular focus on cloud computing.

3.1CiC: Computing in the Cloud

Computing in the Cloud (CiC) is a collaborative cloud computing agreement that Microsoft Corporation and NSF announced in February 2010. Awards were announced April 20, 2011. Microsoft is providing 13 cloud computing research projects with access to Windows Azure, a cloud computing platform that provides on-demand computing and storage to host, scale and manage Web applications on the Internet through Microsoft data centers. These projects will have access to Azure for a two-year period, along with a support team to help researchers quickly integrate cloud technology into their research. We describe a selection of these projects below; see Press Release 11-082 http://www.nsf.gov/news/news_summ.jsp?cntn_id=119248 for full details. (These 13 projects are among the 125 projects mentioned previously.)

A growing spectrum of society-critical, highly sensitive applications is shifting towards cloud computing to benefit from lower costs. These include multiple applications useful in medicine, ranging from treatment to surgical procedures. However, several key issues (such as high availability, secure access, fault tolerance and the preservation of privacy and real-time responsiveness) remain to be addressed. One project will explore the consistency issue of cloud computing applications in large-scale systems, which will contribute towards a scientific foundation for scalable trust in cloud computing.

Today's rapid "data deluge" in the scientific enterprise gives rise to many exciting data mining opportunities. One project explores an alternative data processing architecture, "continuous bulk processing," that will fundamentally improve computing efficiency, reduce costs, and provide enhanced data mining capabilities for cloud computing. A key facet of the approach is to simply update analytics when new data arrives, rather than to recompute them from scratch. The work will explore the ultimate reach of this incremental approach, and determine how users may trade cost for performance for incremental analytics.

Another project will investigate ways to build a cloud storage service under minimal trust assumptions, i.e., without the clients having to assume that the providers will always operate correctly. Trust issues particularly relevant to cloud storage include storage service providers operated by a party other than the data owner, software bugs, correlated manufacturing defects, misconfigured servers, operator error, malicious insiders, bankruptcy, fires, and more.

Accurate forecasting is key to effective utilization of weather-dependent renewable energy sources, such as wind and solar. Weather forecasting is a complex and data-intensive

computational process. One project seeks to develop the Forecast-as-a-Service (FaaS) framework. FaaS will allow new prediction models to use of different types of data from different sources, and hopefully create more accurate forecasts. It will also support on-demand delivery of different types of forecasts at different levels of detail for varying prices, as to accommodate renewable energy users with different needs and varying budgets.

3.2 PROBE: A National Facility for Hosting Clou d Test-Beds

The CSR program in the CNS division of CISE is currently funding a facility called PROBE that hosts Cloud Test-Beds. These test-beds can be utilized by members of the computer systems research community in the United States to carry out their system experiments on large-scale, data intensive, and supercomputing machines. This facility is hosted in the New Mexico Consortium, and is located within the Los Alamos Research Park building, owned by Los Alamos County. The facility is currently supported by NSF, the New Mexico Consortium, the Los Alamos National Laboratory, the University of New Mexico, Carnegie Mellon University, and the University of Utah. PROBE will enable critical experimental research in many computing areas, including reliability, correctness, productivity, and energy efficiency. The facility will have machines and disks in large volumes so that both computing-intensive and disk-intensive experiments can be pursued at scale.

More importantly, PROBE will be created at a fraction of the cost of a typical supercomputing facility, since most of the machines in this facility are ones that were decommissioned from the DOE laboratories. The facility will start with a 1000-node cluster with 2000 processors. Once this cluster is established, a second 1000-node cluster, with 2000 to 3000 processors, will be added. Then, for the foreseeable future, more supercomputing machines can be added annually to the facility as these machines are decommissioned from the DOE laboratories. The annual cost of the facility, including staff, operating costs, and infrastructure, is three million dollars or less, which is indeed a small fraction of the typical operational costs of a large supercomputing center.

3.3 NEBULA: A Future Internet Architecture to Support the Clouds

The NeTS program in the CISE directorate of the NSF has established the Future Internet Architectures (FIA) program to develop four possible architectures for the Internet in the future. One of these four architectures, called NEBULA, is intended primarily to support cloud computing. The NEBULA architecture surrounds a highly available and extensible core network with trustworthy transit and access networks that enable many new forms of communication and computing.

Mobile users of the NEBULA architecture will have fast, secure, anytime/anywhere access to critical services, such as financial transactions and electronic medical services. Local devices will be able to select from a continuum of distributed computing and storage services that are provided by the data centers accessible via the NEBULA architecture.

The NEBULA architecture achieves the three security properties of confidentiality, integrity and availability using a system approach. Specifically, the NEBULA architecture consists of three inter-

related parts: (1) the NEBULA Data Plane that establishes policy-compliant paths and provides both flexible access control and defense against availability and denial-of-service attacks; (2) the NEBULA Control Plane that provides access to application-selectable service and network abstractions such as redundancy, consistency, and policy routing; and (3) the NEBULA Core that redundantly interconnects data centers containing replicated data with ultra-high availability core routers, which are currently being developed in collaboration with Cisco Systems, Inc.

3.4 Cryptography in the Cloud Workshops

Two workshops were help on the topic of cryptography and cloud computing and funded by the Trustworthy Computing program. The first workshop was held August 2nd to 5th, 2009, and focused on the topic of secure cloud computation. This workshop had 65 attendees and 46 presentations during its four days. A second workshop was held August 11th and 1ih, 2010. This workshop discussed verifiable computation in the cloud, i.e., ensuring that the provider will return accurate results for outsourced computation.. Twenty people attended with workshop, which had eleven presentations. Both of these workshops had active participation from companies highly involved in cloud computing, including Microsoft and IBM.

3.5 PI Meeting on the Science of Cloud Computing

A Principal Investigator (PI) meeting on the Science of Cloud Computing was held on March 17-18, 2011 in Arlington, VA. This meeting was attended by 60 Pis from 23 states, 4 industry representatives, and 10 program directors from the CISE directorate of NSF.

The main objective of this meeting was to showcase the Science of Cloud Computing and demonstrate that this science (1) is essential for the economic growth of the United States, (2) is intellectually rich and scientifically deep, (3) has a mature, vibrant, and committed community to develop it.

The meeting concluded by identifying the following critical research areas that need further development and focused funding in the Science and Engineering of Cloud Computing:

1. Cloud Architectures and Systems
2. Network Support for Clouds
3. Data Portability, Consistency, availability, and Management
4. Programming Models for Clouds
5. Fault Masking in the Clouds
6. Cloud Security, Privacy, and Auditing
7. Cloud Debugging, Certification, Diagnosis, and Update
8. Cloud Self-Monitoring and Autonomic Control
9. Cloud Inter-Operability and Standardization
10. Green Clouds

11. Cloud Test-beds

4. Ongoing and Future Initiatives on Cloud Computing

Several efforts are ongoing and several future directions are currently under consideration inCISE. We describe these next.

4.1CISE Core and Cross-cutting Programs

As mentioned previously in section 2, active CISE awards that directly impact cloud computing include 76 managed by the CNS Division, 40 managed by the CCF Division, and 9 managed by the IIS Division. Most of these awards are from programs that are not specifically targeted to cloud computing, with the exception of the CiC program. Titles, amounts, and abstracts of these awards will be made available upon request.

These awards are housed in numerous programs. In CNS, awards come from both core programs: the Computer Systems Research (CSR) program and the Networking Technology and Systems (NeTS) program. The FY 2012 CNS solicitation called out cloud computing as a "highlighted area" for the CSR program. CNS also contributes to several cloud computing awards from the NSF-wide Major Research Instrumentation program. In CCF, awards come from all three of the core programs: Algorithmic Foundations (AF), Communication and Information Foundations (CIF), and Software and Hardware Foundations {SHF); CCF also manages awards made by the CiC program, which was described in section 3. In IIS, awards come from the Information Integration and Informatics (III) program. Many awards come from the cross-division Trustworthy Computing (TC) program.

It is expected that cloud computing will continue to be an active area of research and that these programs will continue to receive and fund proposals in this area.

4.2 Cloud Computing Security Workshop

Since 2009, the Trustworthy Computing Program has sponsored the Cloud Computing Security Workshops, which are held in conjunction with the ACM Conference on Computer and Communications Security. These workshops have been highly successful. They are amongst the most well attended workshops at the conference, and they draw notable participants both from academia and industry. This year's workshop will include keynotes from CA Technologies, Microsoft, and Intel.

4.3 Workshop on the Software Engineering of Cloud Computing

The CSR program and the SHF program are planning a joint workshop in FY 2012 on the Software Engineering of Cloud Computing. This workshop will invite 60 PIs from across the United States to meet and identify the most important problems that need to be solved in order to support development of

software for cloud computing platforms and environments. Topics that will be discussed in this workshop are as follows:

1. Programming Models for Clouds
2. Programming Languages for Clouds
3. Reliable Software for Clouds
4. Secure Software for Clouds
5. Software Components for Clouds
6. Software Debugging for Clouds
7. Software Diagnosis for Clouds
8. Software Certification for Clouds
9. Software Update for Clouds
10. Software Tools for Clouds

4.4 The Science and Engineering of Cloud Computing (SECC}

The CISE directorate is currently considering future directions for cloud computing research under the working title Science and Engineering of Cloud Computing (SECC). SECC is intended to address the important questions of how to design correct and efficient Future Cloud Systems, rather than how to utilize existing cloud systems. Thus, this research represents collaboration between the following five technical areas:

1. Computer Systems, which is supported by the CNS program Computer Systems Research (CSR)
2. Networks, which are supported by the CNS program Networking Technology and Systems (NeTS)
3. Security (including cryptography), which is supported by the CISE program Trustworthy Computing (TC) and by the CCF program Algorithmic Foundations (AF)
4. Computer Architecture and Software, which are supported by the CCF program Software and Hardware Foundations (SHF)
5. Databases and Data-Intensive Computing, which are supported by the IIS program Information Integration and Informatics (III)

It is believed that a focused effort to bring these research communities together will best serve the need expressed by NIST and others to develop and mature the science of cloud computing.

5. Conclusion: Findings and Recommendation

As recognized by both NIST and the America COMPETES Reauthorization Act, cloud computing is an area vital to the economic growth and competitiveness of the nation. The CISE community has responded vigorously to the challenge, as evidenced by the wide range of research efforts underway that are supported by numerous programs throughout the CISE Directorate.

For the most part, the research directions already under exploration by the CISE research community were arrived at naturally, without specific emphasis on cloud computing by NSF. With a few exceptions, these projects already address the critical areas identified at the PI Meeting on the Science of Cloud Computing. The remaining gaps include some areas noted by NIST as barriers to adoption. These critical areas were included in the CISE/CNS Division Core Program Solicitation for FY 2012, which included Cloud Computing as a "Highlighted Area" for its Computer Systems Research (CSR) program. Future solicitations will seek additional research in these areas as appropriate.

NSF anticipates that the CISE directorate will continue its support for Cloud Computing in future years, primarily driven by the CNS Division, with participation by CCF, IIS, the Office of Cyberinfrastructure (OCI), and other Directorates or Offices as appropriate.