The Intel Science and Technology Center for Cloud Computing

Foundations for Future Clouds

White Paper
Intel Labs
ISTC for Cloud Computing

Abstract

The Intel Science and Technology Center for cloud computing (ISTC-CC) is an open community of leading researchers devising critical new underlying technologies for future clouds and cloud applications. Headquartered at Carnegie Mellon University, ISTC-CC includes researchers from Carnegie Mellon, Georgia Tech, Intel, Princeton, and UC-Berkeley. ISTC-CC research explores system architectures, programming models, automation mechanisms, and related technologies to enable dramatic efficiency, ubiquity, and productivity improvements in cloud computing.

Background

Although cloud computing has generated enormous buzz and attention in the industry, the term itself is not always defined with precision. From a user's perspective, cloud computing involves performing a task using someone else's computers and possibly software. The "cloud" aspect refers to resources reached through the network or Internet, which are often drawn as a cloud in technical diagrams. With cloud computing, users access cloud services via many different types of devices, be it a notebook computer, smartphone, tablet, or others. Some examples include using online storage when sharing videos or photos, engaging in social networks, and using online collaboration tools. Datacenters are at the heart of cloud computing and enable the wide variety of online services that business and consumers enjoy today.

Technically speaking, the U.S. National Institute of Standards and Technology (NIST) definition of cloud computing[1], two pages in length, includes this summary:

> Cloud computing is a model for enabling 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.

The importance of cloud computing lies in the fact that this model greatly speeds the deployment of new applications/services, increases the efficiency of operating them, improves agility, and makes it easier to share data and services. It is a paradigm shift for IT application development and delivery, which will enable IT to better handle the explosion of internet users and devices: 2.5 billion people[2] and as many as 15 billion devices[3] will be connected by 2015. Moreover, with a significant increase in content and online services, enterprises are facing ever-increasing demands placed on networks and datacenters that are near capacity, resource-constrained, and increasingly complex. Today's infrastructures will be rapidly exceeded if we do not find a more energy efficient, scalable and cost effective way to handle this growth. For example, Bain estimates about $2 trillion will be spent on IT deployment and operations thru 2015[4] without a more simplified infrastructure.

Huge potential benefits

Innovation in this field is important for both consumers and the enterprise. Cloud computing connects people to essential information, powers the mobile & social media revolutions, enables new online services, and improves the way we live and work.

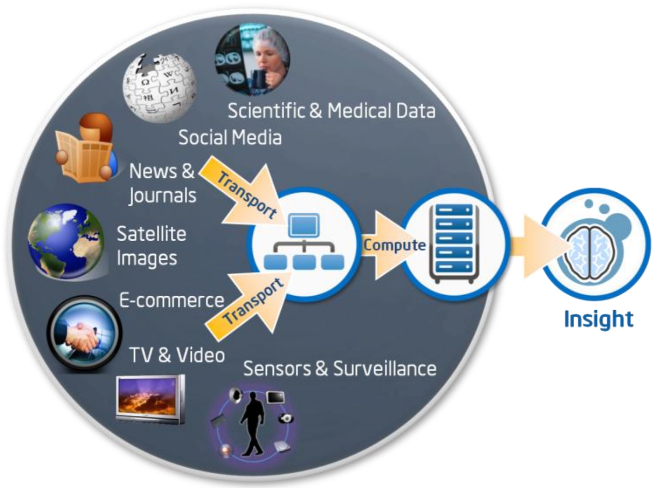

Figure 1 New insight will be gained from combining, computing, and drawing meaning from the exploding wealth of data in the cloud.

Analyzing massive online datasets available in the cloud will lead to new scientific and medical discoveries, the ability to monitor and react to real-time events like emergencies, and other breakthroughs. A few example applications include:

- Using cloud processing to decode brainwave patterns and create new user interfaces.
- Using high-performance processors in the cloud to run and stream high-end games with advanced graphics to any and all of your devices.

[1] http://www.nist.gov/itl/cloud/upload/cloud-def-v15.pdf
[2] IDC "The Internet Reaches Late Adolescence" Dec 2009, extrapolation by Intel for 2015
[3] Intel Embedded & Communications Group forecast "Worldwide Device Estimates Year 2020 - Intel One Smart Network Work"
[4] Intel commissioned analysis with Bain Consulting based on: IDC "Server Workloads Forecast" 2009, AMI US SB/MB Overview and Market Opportunity Assessment 2006 Gartner "WW IT Services Forecast" 2009; Cowen & Co "VMware" (July 2009); IDC 2009 "Optimizing Infrastructure and Service Management in Tough Economic Times", analyst reports, IDC storage & networking reports, Gartner 2009 WW IT Services Forecast, BLS, Computer Economics, Principled Technology, expert interviews, Bain analysis.

- Emergency response training by creating massive, multiplayer simulations in which people play the roles of victims and responders.
- Using computation to automatically identify cancerous cells, such as by capturing images of cells, comparing to massive databases of existing images, and identifying likely occurrences.

For businesses, the most commonly noted benefits relate to efficiency, agility and speed of provisioning resources, and lower costs. With cloud computing, multiple customers share providers' offerings instead of each providing for itself. Efficiency improvements come from increased utilization of resources, economies of scale, increased IT staff specialization, and low entry cost for new customers. For example, Intel IT recently developed a cloud-based grid for silicon microprocessor designers. This design grid created a pool of compute resources, including tens of thousands of servers, located across multiple sites. By sharing these resources across sites, Intel IT significantly increased design compute server utilization. This, together with proactive server refresh, has reduced the need to add data center capacity, which will save the company an estimated $200 million.[5]

Significant Challenges

Despite the excitement and progress, cloud computing is still in an early stage. Companies are creating public cloud services and private clouds, but each does so differently without a consensus regarding compatible interfaces that enable interoperability among clouds, mechanisms, or policies for cloud services. Existing services are changing rapidly as experience is gained. Similarly, the surface has barely been scratched regarding applications and uses of cloud computing. The security of personal and confidential data in the cloud is also a growing concern. Nonetheless, the great potential benefits of cloud computing, combined with the momentum driving it, give it a bright future.

The Intel Science and Technology Center for Cloud Computing (ISTC-CC) was created to help address these challenges and more fully realize the potential of cloud computing. The center is a research collaboration between Intel Labs and cloud computing experts from outstanding academic institutions including Carnegie Mellon University, Princeton, Georgia Tech, and UC Berkeley. Carnegie Mellon will be the hub of the center, coordinating these activities. The ISTC-CC will be co-led by Carnegie Mellon's Professor Greg Ganger and Intel's Dr. Phil Gibbons. ISTC-CC will create a vibrant cloud computing research community, fostering and embracing open collaboration with other researchers, including at other universities and in other ISTCs.

Research Vision

Beyond 2015

Intel is advocating a vision for the future of cloud computing called 'Cloud 2015.' This vision foretells of clouds with three key characteristics: 'Federated' such that enterprises can share data and services seamlessly and securely across internal and external clouds; 'Automated' so that resources are dynamically allocated to maximize efficiency, quality of service, and productivity; and 'Client-aware' meaning cloud-based applications are able to dynamically sense and take advantage of the capabilities of the end-point device to optimize application delivery in a secure fashion while delivering a great user experience. To achieve this vision, Intel is focused on delivering leading technologies and collaborating with the industry ecosystem to enable solutions that enhance security, improve efficiency, and enable more simplified clouds.

The ISTC-CC will expand on Intel's cloud vision with insight from top academic researchers, and includes advanced research that we expect will extend and improve upon the elements above such as building in advanced automation, real-world context awareness from client sensors, and large-scale specialization for better efficiency. In addition, we are looking well beyond 2015 to explore technology that we predict will be important for the cloud in the future, including more widespread and better analytics to gain insight from massive amounts of online data and making the cloud more decentralized and location-aware by extending cloud capabilities to the network edge and client devices.

Our vision for ISTC-CC research focuses on game-changing underlying technologies needed for cloud computing benefits to become pervasive within the next 10 years. We will also take advantage of connections with other ISTCs for Secure, Embedded and Visual Computing for important complementary research topics, including improving cloud security and inventing new cloud applications.

[5] Intel, "An Enterprise Private Cloud Architecture and Implementation Roadmap," June 2010, www.intel.com/IT.

Future cloud computing will incorporate heterogeneous mixes of specialized platforms utilizing a variety of emerging technologies. It will use and be used for big data analytics on stored and live data feeds as well as encompass billions of edge devices through new paradigms for mash-ups of clients and cloud. It will also rely on significant advances in automation to realize the desired efficiency and productivity.

The ISTC-CC research agenda is organized into four inter-related research pillars architected to create a strong foundation for cloud computing of the future:

Figure 2 ISTC-CC pillars providing foundation for cloud computing of the future. A broad collection of applications and computing activities will rely on cloud computing infrastructures

Specialization

More than 2% of electricity in the U.S. is consumed by datacenters[6]. Without innovation to increase efficiency, the equivalent of 45 new coal power plants may be needed by 2015[7] to fuel the growing cloud deployments.

This pillar explores the use of specialization as a primary means for order of magnitude improvements in energy consumption. Contrary to the common practice of striving for homogeneous cloud deployments, we envision clouds that embrace heterogeneity, purposely including mixes of different platforms specialized for different classes of applications, such as designs based on emerging technologies like non-volatile memory and specialized cores. For instance, when a consumer logs into a website, the server should be optimized for simple transactions. When a scientist is running a web-based visualization using advanced digital effects like ray tracing, the server should accelerate software execution with specialized hardware such as Intel's Many Integrated Core (MIC) architecture.

Automation

Operational costs are a sizable and growing portion of datacenter total cost of ownership (TCO). A recent Intel analysis of a large-scale Internet datacenter (see Figure 3 below) showed of that human administration, down-time induced losses, energy use, etc. accounted for roughly half of TCO. Advanced automation will be one of the keys to help reduce these costs and improve overall efficiency for 2015 and beyond. The scale, diversity, and unpredictability of cloud workloads increase both the need for, and the challenge of, automation. This pillar addresses cloud's particular automation challenges, focusing on order of magnitude efficiency gains from smart resource allocation/scheduling (including automated selection among specialized platforms) and greatly improved problem diagnosis capabilities. For example, when an application service's performance is lower than expected in a cloud setting, new tools are needed to pinpoint the root cause(s) among the many possibilities, ranging from application shortcomings, issues with software components on which the application relies, and cloud infrastructure issues (e.g., insufficiently managed interference from other tenant activities).

[6] EPA Report to Congress on Server and Data Center Energy Efficiency; August 2, 2007

[7] Power savings calculated based on projected performance improvements from Intel roadmap while keeping power / system flat. Moore's Law drives ~2x perf / 18 months. At 5 years, that equals 10X. We assume that compared to 2010, we're saving 9X (i.e, the 10x less the 1X for what you'll need). It assumes we keep power per system constant at 200W. We assume we'll need 16M servers in 2015 (based on market model) – that means we save 16M x 9X x 200W (average system power) x 1.6 PUE = ~45GW. The estimated power/coal plant is 1Gw 45GW = ~ 45 coal plants needed.

Big Data

The world is constantly generating a massive wealth of information accumulating across the internet, from video and tweets to business and scientific data. IDC predicts a 75X growth within the next 10 years in number of files enterprises deal with[8] along with an explosion in data types from structured (e.g., relational database) to unstructured (video, photos, emails, documents, information from sensors and more). The ability to draw meaning from these massive datasets will lead to new scientific breakthroughs, business insights, and a better understanding of our increasingly complex world.

The term "Big Data" refers to the large and often dynamically growing datasets on which the new data-driven approach to knowledge creation, sometimes called "data-intensive computing" or DISC, relies. Generally, DISC applies various forms of statistical machine learning to induce models and extract insights from Big Data. Depending on the application, Big Data may include datasets ranging from hundreds of gigabytes in size to many terabytes or even petabyes. This pillar addresses the critical need for cloud computing to extend beyond early, first-generation big data usage (primarily, search) to efficiently and effectively support Big Data analytics, including new high-productivity frameworks for advanced machine learning algorithms and for coping with the continuous ingest, integration, and exploitation of live data feeds such as video streams or social networks. The applications of this capability are diverse and critical to our futures – from understanding widespread scientific trends across nearly every field to diagnosing medical problems to solving crimes by monitoring video and social media to tracking and responding to widespread social events such as epidemics and natural disasters.

To the Edge

Future cloud computing will extend beyond centralized (back-end) resources by encompassing billions of clients and edge devices (local network components, local cloud servers, and mobile user devices) and taking advantage of distributed servers from mega-data centers to the edge. Sensors, actuators, and "context" provided by such devices, even those with limited capabilities, will be among the most valuable content resources in the cloud. This pillar explores new frameworks for device/cloud cooperation with an enhanced edge that can efficiently and effectively exploit this "physical world" content in the cloud as well as enable localized, cloud-assisted client computations. That is, applications whose execution spans client devices, edge-local cloud resources, and core cloud resources. Research in this area could lead to technologies such as a digital personal handler. Imagine a device wired into your glasses that sees what you see, constantly pulls data from the cloud, and whispers hints to you during your day – telling you who people are, where you can buy that cool thing you just saw, or how to adjust your plans when something new comes up.

Additional Collaborative Research

Cloud Security

Security is a subject of critical importance to cloud adoption. A recent IDC survey revealed that over 70%[9] of the IT industry is concerned about security in the public cloud. Intel's ISTC program has put a special emphasis on security by launching the ISTC for Secure Computing (ISTC-SC) focused on this subject. This center, headquartered at UC Berkeley, includes a broad security research agenda[10] spanning both cloud and client. Both of these ISTCs involve both Carnegie Mellon and UC Berkeley, and they will work collaboratively on problems of cloud security.

A key example of this research relates to the protection of personal data as it traverses the cloud. Data increasingly flows through complex distributed systems and is stored in various forms at a variety of locations. Thus, it is essential to develop consistent and reliable protection for user data, no matter where it is retained and used. The ISTC for Secure Computing takes a multi-pronged approach including developing a means for attaching security policies to data, privacy-preserving services, and, for legacy applications, wrapper techniques for information-flow tracking.

Cloud Testbeds

ISTC-CC will participate in maintaining several cloud testbed infrastructures, addressing two goals: fostering a robust cloud research community and gathering invaluable information regarding cloud demands and

[8] IDC Digital Universe Study, June 2011

[9] 1. IDC Market Analysis, January 2010.

[10] http://istcsc.cs.berkeley.edu/docs/ISTC-SC-Whitepaper.pdf

deployments. They also will enable experiments in infrastructure, tools, and platform specialization.

One example is the OpenCirrus[11] testbed, which includes existing deployments at Intel, Carnegie Mellon and Georgia Tech, and other leading universities and which also creates excellent opportunities for cloud federation explorations. OpenCirrus is an open cloud computing testbed designed to foster research in both cloud computing applications and cloud provider systems. Started in 2008 by Intel, HP and Yahoo!, Open Cirrus federates a heterogeneous collection of resources distributed across data centers around the world. Currently, there are 15 sites worldwide, and each site consists of a cluster with at least 1000 cores and associated storage. Intel's OpenCirrus cluster alone consists of 1508 cores across 210 servers, supporting 40+ projects and 100+ users.

Cloud Applications

These testbeds will also help foster collaborations with cloud application researchers by providing them with access to cloud resources. For example, we intend to collaborate with the ISTC for Embedded Computing (also headquartered at Carnegie Mellon), the ISTC for Visual Computing (headquartered at Stanford University), as well as other academic collaborators. This will also provide ISTC-CC researchers with invaluable data and case studies regarding workloads, resource utilization, failures, and operational aspects.

More information other ISTC research is available via www.intel.com/go/istc.

Pillar 1: Specialization

Driving greater efficiency is a significant global challenge for cloud datacenters. Current approaches to cloud deployment, especially for increasingly popular private clouds, follow traditional data center practices of identifying a single server architecture and avoiding heterogeneity as much as possible. IT staff have long followed such practices to reduce administration complexity—homogeneity yields uniformity, simplifying many aspects of maintenance, such as load balancing, inventory, diagnosis, repair, and so on. Current best practice tries to find a configuration that is suitable for all potential uses.

Unfortunately, there is no single server configuration that is best, or close to best, for all applications. Some applications are computation-heavy, needing powerful CPUs and little I/O bandwidth, while others are I/O-bound and involve large amounts of random I/O requests. Some are memory-limited, while others process data in streams (from storage or over the network) with little need for RAM. And, some may have characteristics that can exploit particular hardware assists, such as GPUs, encryption accelerators, and so on. A multi-purpose cloud could easily see a mix of all of these varied application types, and a lowest-common-denominator type configuration will fall far short of best-case efficiency.

We believe that specialization is crucial to achieving the best efficiency—in computer systems, as in any large-scale system (including society), specialization is fundamental to efficiency. Future cloud computing infrastructures will benefit from this concept, purposefully including mixes of different platforms specialized for different classes of applications. Instead of using a single platform configuration to serve all applications, each application (and/or application phase, and/or application component) can be run on available servers that most closely match its particular characteristics. We believe that such an approach can provide order-of-magnitude efficiency gains, where appropriate specialization is applied, without yielding the economies of scale and elastic resource allocation promised by cloud computing.

Additional platforms under consideration include lightweight nodes (such as nodes that use Intel® Atom processors), heterogeneous many-core architectures, and CPUs with integrated graphics, with varied memory, interconnect and storage configurations/technologies.

Realizing this vision will require a number of inter-related research activities:

- Understanding important application classes, the trade-offs between them, and formulating specializations to optimize performance.
- Exploring the impact of new technologies like non-volatile memory (NAND flash, phase change memory, etc.).
- Creating algorithms and frameworks for exploiting such specializations.
- Programming applications so that they are adaptable to different platform characteristics, to maximize the

[11] www.opencirrus.org

benefits of specialization within clouds regardless of the platforms they offer.

In addition, the heterogeneity inherent to this vision will also require new automation approaches.

Pillar 2: Automation

As computer complexity has grown and system costs have shrunk, operational costs have become a significant factor in the total cost of ownership. Moreover, cloud computing raises the stakes, making the challenges tougher while simultaneously promising benefits that can only be achieved if those challenges are met.

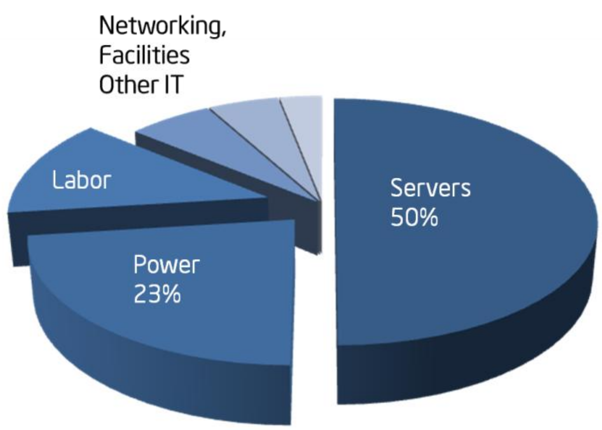

Figure 3 Intel analysis of typical large-scale internet datacenter total cost of ownership over 3 years shows operational costs equal the cost of servers.

Operational costs include human administration, downtime-induced losses, and energy usage. Administration expenses arise from the broad collection of management tasks, including planning and deployment, data protection, problem diagnosis and repair, performance tuning, software upgrades, and so on. Most of these become more difficult with cloud computing, as the scale increases, the workloads run on a given infrastructure become more varied and opaque, workloads mix more (inviting interference), and pre-knowledge of user demands becomes rare rather than expected. And, of course, our introduction of specialization (Pillar 1) aims to take advantage of platforms tailored to particular workloads.

Automation is the key to driving down operational costs. With effective automation, any given IT staff can manage much larger infrastructures. Automation can also reduce losses related to downtime, both by eliminating failures induced by human error (the largest source of failures) and by reducing diagnosis and recovery times, increasing availability. Automation can significantly improve energy efficiency, both by ensuring the right (specialized) platform is used for each application, by improving server utilization, and by actively powering down hardware when it is not needed.

Within this broad pillar, ISTC-CC research will tackle key automation challenges related to efficiency, productivity and robustness, with three primary focus areas:

- Resource scheduling and task placement: devising mechanisms and policies for maximizing several goals including energy efficiency, interference avoidance, and data availability and locality. Such scheduling must accommodate diverse mixes of workloads as well as specialized computing platforms.
- Devising automated tools for software upgrade management, runtime correctness checking, and programmer productivity that are sufficiently low overhead to be used with production code at scale.
- Problem diagnosis: exploring new techniques for diagnosing problems effectively given the anticipated scale and complexity increases coming with future cloud computing.

Pillar 3: Big Data

Data-intensive scalable computing (DISC) refers to a rapidly growing style of computing characterized by its reliance on large and often dynamically growing datasets ("BigData"). With massive amounts of data arising from such diverse sources as telescope imagery, medical records, online transaction records, checkout stands and web pages, many researchers and practitioners are discovering that statistical models extracted from data collections promise major advances in science, health care, business efficiencies, and information access. In fact, in domain after domain, statistical approaches are quickly bypassing expertise-based approaches in terms of efficacy and robustness.

The shift toward DISC and Big Data analytics pervades large-scale computer usage, from the sciences (e.g., genome sequencing) to business intelligence (e.g., workflow optimization) to data warehousing (e.g., recommendation systems) to medicine (e.g., diagnosis) to Internet services (e.g., social network analysis) and so on. Based on this shift, and their resource demands relative to more traditional activities, we expect DISC and Big

Data activities to eventually dominate future cloud computing.

We envision future cloud computing infrastructures that efficiently and effectively support DISC analytics on Big Data. This requires programming and execution frameworks that provide efficiency to programmers (in terms of effort to construct and run analytics activities) and the infrastructure (in terms of resources required for given work). In addition to static data corpuses, some analytics will focus partially or entirely on live data feeds (e.g., video or social networks), involving the continuous ingest, integration, and exploitation of new observation data.

ISTC-CC research will devise new frameworks for supporting DISC analytics of Big Data in future cloud computing infrastructures. Three particular areas of focus will be:

▪ Understanding DISC applications, creating classifications and benchmarks to represent them, and providing support for programmers building them.

▪ Frameworks that more effectively accommodate the advanced machine learning algorithms and interactive processing that will characterize much of next generation DISC analytics.

▪ Cloud databases for huge, distributed data corpuses supporting efficient processing and adaptive use of indices. This focus includes supporting datasets that are continuously updated by live feeds, requiring efficient ingest, appropriate consistency models, and use of incremental results.

Note that these efforts each involve aspects of Automation, and that Big Data applications represent one or more classes for which Specialization is likely warranted. The aspects related to live data feeds, which often originate from client devices and social media applications, lead us into the next pillar.

Pillar 4: To the Edge

Future cloud computing will be a combination of public and private clouds, or hybrid clouds, but will also extend beyond large datacenters that power cloud computing to include billions of clients and edge devices. This includes networking components in select locations and mobile devices closely associated with their users that will be directly involved in many “cloud” activities. These devices will not only use remote cloud resources, as with today’s offerings, but they will also contribute to them. Although they offer limited resources of their own, edge devices do serve as bridges to the physical world with sensors, actuators, and “context” that would not otherwise be available. Such physical-world resources and content will be among the most valuable in the cloud.

Effective cloud computing support for edge devices must actively consider location as a first-class and non-fungible property. Location becomes important in several ways. First, sensor data (e.g., video) should be understood in the context of the location (and time, etc.) at which it was captured; this is particularly relevant for applications that seek to pool sensor data from multiple edge devices at a common location. Second, many cloud applications used with edge devices will be interactive in nature, making connectivity and latency critical issues; devices do not always have good connectivity to wide-area networks and communication over long distances increases latency.

We envision future cloud computing infrastructures that adaptively and agilely distribute functionality among core cloud resources (i.e., backend data centers), edge-local cloud resources (e.g., servers in coffee shops, sports arenas, campus buildings, waiting rooms, hotel lobbies, etc.),and edge devices (e.g., mobile handhelds, tablets, netbooks, laptops, and wearables). This requires programming and execution frameworks that allow resource-intensive software components to run in any of these locations, based on location, connectivity, and resource availability. It also requires the ability to rapidly combine information captured at one or more edge devices with other such information and core resources (including data repositories) without losing critical location context.

ISTC-CC research will devise new frameworks for edge/cloud cooperation. Three focus areas will be:

▪ Enabling and effectively supporting applications whose execution spans client devices, edge-local cloud resources, and core cloud resources, as discussed above.

▪ Addressing edge connectivity issues by creating new ways to mitigate reliance on expensive and robust Internet uplinks for clients.

▪Exploring edge architectures, such as resource-poor edge connection points vs. more capable edge-local servers, and platforms for supporting cloud-at-the edge applications.

Conclusion: Delivering Results

The primary ISTC-CC deliverables will be new concepts, algorithms, and system designs resulting from the research activities described in this proposal. In keeping with the open community spirit of the ISTC model, these deliverables will be reported in the open literature, via publication in top conferences, in technical reports, and in other forums. We also plan to share software prototypes under open source licenses wherever possible.

We envision a future where the cloud is ubiquitous, with pervasive use of storage and computing done via cloud computing resources. Client edge devices will interface to the physical world and may perform some work themselves; this work will complement larger computations, such as large-scale data analytics, occurring in the cloud in various forms. Data will be secured and, of course, this envisioned future would enjoy all of the potential efficiency and productivity benefits we expect from the cloud.

Printed in USA   Please Recycle   ISTC_CC-2011_001US