# The future of cloud computing: an HP Labs perspective

Prith Banerjee, Cullen Bash, Rich Friedrich, Patrick Goldsack, Bernardo A. Huberman, John Manley, Lueny Morell, Chandrakant Patel, Partha Ranganathan, Martina Trucco, Alistair Veitch

HP Laboratories
HPL-2010-192

**Keyword(s):**

cloud computing, data center, energy management, systems, architecture, storage system, social computing

**Abstract:**

The key elements required to enable an "Everything-as-a-Service" economy include technologies for servers, storage, networking and IT management necessary to deliver Infrastructure as a Service, technologies for a shared cloud infrastructure that provides enterprise-grade security, scalability and quality of service for delivering Platforms as a Service, and novel consumer and enterprise services for the cloud delivered as Software as a Service.

External Posting Date: December 21, 2010 [Fulltext] Approved for External Publication
Internal Posting Date: December 21, 2010 [Fulltext]
To be published in IEEE Computer.

# THE FUTURE OF CLOUD COMPUTING: AN HP LABS PERSPECTIVE

Prith Banerjee, Cullen Bash, Rich Friedrich, Patrick Goldsack, Bernardo A. Huberman, John Manley, Lueny Morell, Chandrakant Patel, Partha Ranganathan, Martina Trucco, Alistair Veitch, *Hewlett Packard Laboratories*

---

The key elements required to enable an "Everything-as-a-Service" economy include technologies for servers, storage, networking and IT management necessary to deliver *Infrastructure as a Service,* technologies for a shared cloud infrastructure that provides enterprise-grade security, scalability and quality of service for delivering *Platforms as a Service,* and novel consumer and enterprise services for the cloud delivered as *Software as a Service*.

As one of the fastest growing segments of the IT industry, cloud computing has transformed from a promising business concept to a working reality for both enterprises and small and medium businesses. The catalysts that are making this possible are both business- and technology-driven. Cloud services—highly scalable and elastic technology-enabled services, delivered and consumed over the Internet on a one-to-many, as-needed, pay-per-use basis—are enabling businesses to gain flexibility and reduce costs.

Most enterprises are creating a hybrid service portfolio comprising services from many sources – traditional services, private cloud services, and public cloud services. It is critical that organizations align the right model to the right service. And this means they have to build, consume and manage the right cloud services in a secure and economical manner.

HP Labs, HP's central research organization, is chartered with conducting both fundamental and applied research targeted at creating the next breakthroughs in Information Technology [1]. Employing a multi-disciplinary approach, we believe that the greatest opportunities for innovation come from working together in large teams, with customers and partners, on an ambitious shared goal. The HP Labs' research agenda is multi-faceted. **Cloud** and **Security** research is focused on delivering a secure application and computing end-state of *Everything-as-a-Service*: billions of users, securely accessing millions of services, through thousands of service providers, over millions of servers, processing exabytes of data, securely delivered through terabytes of network traffic. **Services** research is focused on enabling corporations to implement new business models utilizing a global ecosystem of IT services delivered over the cloud and developing the next generation of analytical technologies and solutions that will bring personalized experiences to individuals and unprecedented operational efficiencies to the enterprise. **Information Analytics** research enables near real-time business intelligence with robust, scalable data management, data-intensive analytics of structured and unstructured information, and automatically selected, gathered and transformed to fulfill the demand for timely delivery to enterprise customers of information described in the language of business. **Intelligent Infrastructure** research focuses on the technologies (compute, storage, photonics, sensors, etc.) to build and manage next generation infrastructure fabrics for Enterprise and Consumer IT, Cloud, and embedded solutions with dramatically improved scale, performance, reliability and cost of ownership. **Networks** research creates the next generation of network architectures and technologies which provide predictable, high-quality and power-efficient networking while reducing management complexity. **Mobile and Immersive Experience** research creates compelling experiences that fundamentally change how people communicate, collaborate, socialize and entertain with multimedia technologies to deliver interactive, mobile, and immersive audio-visual experiences; develop natural and intuitive forms of interaction between people and technology; and create the next generation of 2-D and 3-D display technologies and information surfaces for mobile and immersive environments. **Commercial Digital Print** research accelerates the analog to digital transformation of commercial print by enabling anyone to be a professional author of personalized content with automated workflows and print engines having breakthrough advances in price/performance, in an ecosystem of secure lifecycle document management.

This paper highlights some of HP Labs' recent cloud computing research. It is a broad and deep agenda with a decade-long history of seminal results. Three key elements required to enable an "Everything-as-a-Service" economy will be described:

- The technologies for servers, storage, networking and IT management necessary to deliver *Infrastructure as a Service;*
- The technologies for a shared cloud infrastructure that provides enterprise-grade security, scalability and quality of service for delivering *Platforms as a Service*;
- Novel consumer and enterprise services for the cloud delivered as *Software as a Service*.

## INFRASTRUCTURE AS A SERVICE

The two fundamental types of services that can be offered under the broad description of Infrastructure as a Service (IaaS) are compute and storage. Compute service typically takes the form of offering virtual machine instances, upon which the customer can install and run whatever software they choose. Storage is more complicated due to the variety of ways in which storage services are offered (e.g. block devices, file systems, databases, each of which have multiple variations.) to applications.

### Scalable Storage for the Cloud

At HP Labs over the past several years, research has been conducted in architecting and prototyping scalable storage systems for the Cloud. Our goals for this research are ambitious: capacity in excess of 1 Exabyte distributed over many geographically distributed data centers, while achieving an availability of 99.999±%, zero data loss, and low cost.

The core problems associated with building any cloud storage system are reliability, scalability and cost-effectiveness. HP Labs believes that reliability (which encompasses both the durability of the data, and its availability for access) is the primary property that users desire from such a system. Whenever a cloud service of any popularity becomes unavailable, or loses data, it quickly becomes front page news [2], as potentially millions of users and thousands of businesses depend on it. This is especially problematic when we consider that to meet the scale and cost-effectiveness goals, such systems must be built from clusters of commodity servers and disk drives, often connected via Ethernet, and spread over multiple data centers. There are a large number of failure scenarios to deal with in such environments, from individual disk and node failures, to network infrastructure outages, power distribution outages, and even disaster scenarios in which an entire data center might effectively be unusable. To cope with these, and ensure overall system reliability, requires enough redundancy in the data to enable recovery from a substantial number of these failures.

A class of design tradeoffs that distributed systems developers must address was popularized by Eric Brewer as the CAP theorem [3]. This essentially states that any distributed system can only offer any 2 of 3 properties: *consistency* (will the result of reads always be consistent with the most recent write?), *availability* (can I access my data?) and *partition tolerance* (can my system tolerate being split into multiple parts, perhaps as a result of network failure?). Because availability is so important, many cloud systems optimize for this, even in the presence of network partitions, and give up on consistency in various failure scenarios. A substantial part of HP Labs' research has been on defining and quantifying the *consistability* of these types of storage systems [4, 5, 6]. Consistability tries to capture the fact that a storage system provides different kinds of consistency under different operating conditions. Currently, developers of cloud-based applications have very little idea what semantics the underlying storage system offers in various corner cases.

The traditional approach to achieving reliability is to replicate data, often many times. However, this has the significant disadvantage that it drives up the underlying cost of the system. Instead, HP Labs has developed a system that uses erasure coding, where data is encoded into $n = k + m$ fragments, such that any $k$ fragments can be used to recover the object, as shown in Figure 1. The use of erasure coding complicates the implementation of such a system, but results in both much greater data reliability at reduced cost. In fact, users of the system can specify different encodings to achieve different cost/capacity/reliability tradeoffs. For example, temporary, easily recreated data could be stored with minimal redundancy and cost, or archival data could be very widely dispersed to ensure complete reliability.

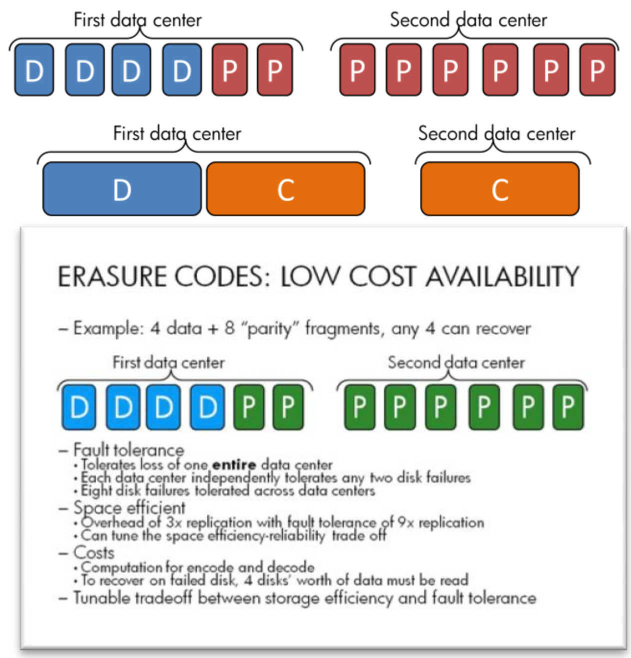


Figure 1: Erasure coding and replication. Erasure coding takes an object, and splits it up into fragments, calculating redundant "parity" fragments, such that a subset of the fragments can be used to recreate the original data. The top part of the figure shows a 4-of-12 erasure code, in which any 4 of the 12 fragments (D for original data, P for parity) can be used to recreate the data. Spread across 2 data centers, this encoding can tolerate the loss of any 8 fragments, or an entire data center and 2 fragments in the second. Three-way replication (shown in the bottom part of the figure) can tolerate far fewer failures, while using the same amount of capacity overall. However, replication is far simpler to implement and reason about in building a distributed system.

Our initial approach has been focused on building a key-value store. A system of this type stores values (an arbitrary string of bytes, representing the object to be stored), associated with a key (a relatively small byte string). Core operations for such a system are *put(key, value)* (associate a value with a key), *get(key)* (retrieve a value associated with a key) *enumerate()* (retrieve key's used) and *delete* (remove an object and it's key from the system). In practical use API's are considerably richer than this, including notions of users, security etc., but these form the base set. The high level architecture of the system developed is shown in Figure 2. *Clients* are the users of the system, who submit requests to *proxies*, who execute those requests, and interface with the remainder of the system. *Key Lookup Servers*, provide metadata lookup, translating keys to the locations of the fragments of data, which are stored in Fragment Servers (FS). This enables scalability, as both types of nodes can be added as needed to support more capacity and/or performance for the system.

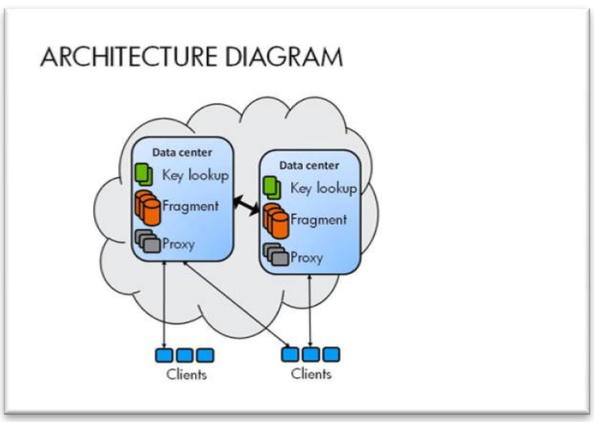


Figure 2: System architecture for storage management in Infrastructure as a Service

A working prototype of the key-value system has been built and tested extensively. A framework was also developed for the implementation that allows for testing code both in simulated environments and on real hardware. It has been used under a wide variety of different failure conditions, including partitions, 50% random packet loss, disk failures, node failures, rack failures, etc., and has verified that availability has been maintained throughout. To ensure that the system behaves as expected under real-world conditions test instances across multiple continents have been executed.

## Energy-efficient Compute Infrastructure for the Cloud

The computing infrastructure is undergoing an unprecedented revolution driven by the emergence of large-scale data centers. These data centers present unique challenges for server design. Specifically, these infrastructures have millions of users running on thousands of servers, making the ability to scale-out server configurations an important design requirement. Data center infrastructure – including power and cooling – can be one of the largest capital and operating expenses for cloud companies motivating them to focus on the sweet spot of commodity pricing and energy efficiency.

Furthermore, innovations in the software stack (including approaches for scalable storage like those discussed above) allow for exploration of novel approaches targeted at cloud datacenters. The Exascale Datacenter research at HP Labs has been addressing these challenges and opportunities. Workloads and metrics have been developed and used to build and analyze new system architectures optimized for the cloud.

First, a near-term architecture optimized for the cloud – called *microblades* and *megaservers* [7] – has been created. This new design involves modular cost-effective server blocks ("microblades") to build large, powerful computing environments ("megaservers"). The design (Figure 3) includes several novel features: disaggregated blade servers including compute blades that use embedded/mobile processors and memory blades that use flash-based non-volatile memory and provide a common second-level memory pool and a novel physical packaging for optimized power and cooling. Evaluation results (Figure 4) shows that this approach has the potential for dramatic improvements compared to the state-of-the-art, improving energy efficiency on average by factors of 4 to 6.

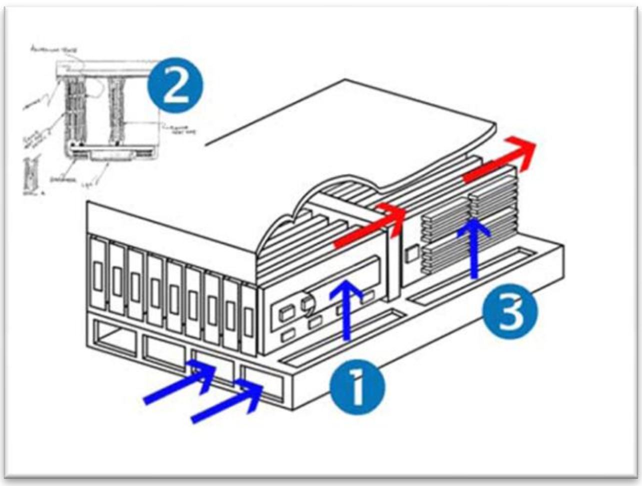


Figure 3: Physical system design of energy-efficient microblade and megaserver compute infrastructure

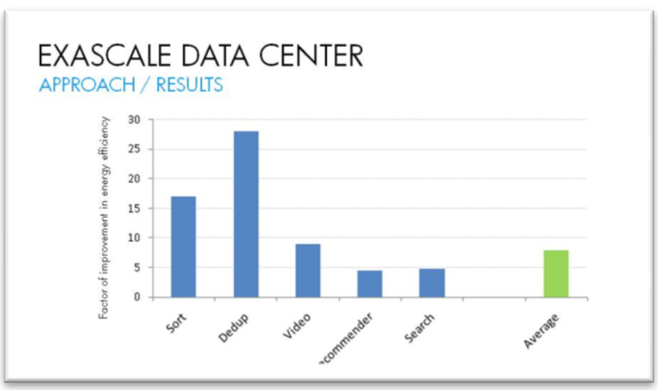


Figure 4: Example energy efficiency improvements

The second example [8] is a novel cross-layer power management solution targeting energy efficiency constraints for cloud datacenters. Currently, the emergent behavior from the collection of individual optimizations of current systems designs may or may not be globally optimal, or even stable, or correct. A key need, therefore, is a carefully designed coordination framework that is flexible and extensible and minimizes the need for global information exchange and central arbitration. As part of a collaboration effort between computer scientists, thermo-mechanical engineers, and control engineering experts, HP Labs developed a novel coordination solution that addresses this need. The design is based on carefully connecting and overloading the abstractions in current implementations to allow the individual controllers to learn and react to the effect of other controllers the same way they would respond to changes in workload demand variations. This enables formal mathematical analysis of stability, and provides flexibility to dynamic changes in the controllers and system environments. A coordination architecture for five individual solutions using different techniques and actuators to optimize for different goals at different system levels across hardware and software has been demonstrated, and HP Labs has shown that this solution can provide significant advantages to existing state-of-the-art.

Beyond these two examples, we are continuing to look ahead to the next-generation of data-centric workloads and new infrastructures to support those. Specifically, emerging non-volatile memories like *memristors*, in combination with advances in photonics and multicore processing, offer intriguing opportunities for new system designs (e.g., *nanostores* [11]) that may offer significantly better performance and energy efficiency. These improvements in future system architectures will enable applications previously not possible in the cloud, for more sophisticated insight generation across larger diverse multiple data sources.

## The Balance of Power: Sustainable Cloud Data Centers

With respect to sustainability, cloud services have the opportunity to change existing business models and deliver a net positive impact by reducing the consumption of the global pool of available energy. However, to reach the desired price point where billions of users accessing millions of services will be feasible—especially in growth economies, where Internet access is desired at approximately US $1 per month—the *total cost-of-ownership* (TCO) of the physical infrastructure that supports the cloud will need to be revisited. There has been progress in reducing the cost of access devices [12], but the cost to access services still needs to be addressed. In this regard, without addressing the cost of data centers—the foundation for services to the masses—scaling to billions of users is not possible.

With respect to data centers, prior work at HP Labs has shown that a significant fraction of the TCO comes from the recurring energy consumed in the operation of the data center, and from the burdened capital expenditures associated with the supporting physical infrastructure. The burdened cost of power and cooling, inclusive of redundancy, is estimated to be 25% to 30% of the total cost of ownership in typical enterprise data centers [13]. These power and cooling infrastructure costs may match, or even exceed, the cost of the IT hardware within the data center. Thus, including the cost of IT hardware, over half of the TCO in a typical data center is associated with design and management of the physical infrastructure. For cloud services providers, with thinner layers of software and licensing costs, the physical infrastructure could be responsible for as much as 75% of the TCO.

The key principle of HP Labs' systematic sustainability framework is supply and demand side management [14]. On the supply side, design of cloud data centers and services ought to focus on: (1) minimizing available energy required to extract, manufacture, mitigate-waste, transport, operate and reclaim components; and (2) design and management using local sources of available energy to minimize the destruction of available energy in transmission and distribution, e.g., dissipation in transmission; and, take advantage of available energy in the waste streams, e.g., exhaust heat from turbine. On the demand side, design of cloud data centers and services ought to focus on minimizing the consumption of available energy by provisioning resources based on the needs of the user by using flexible building blocks, pervasive sensing, communications, knowledge discovery and policy based control. Results to date, illustrated in Figure 5, show a 41% reduction in the lifecycle energy footprint and a 48% reduction in total cost of ownership.

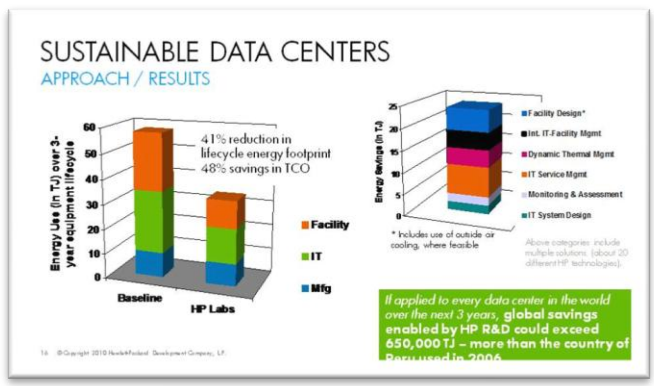


Figure 5: 3-year lifecycle energy improvements

Sustainable cloud services, given this framework, are built on provisioning supply and demand side resources based on the needs of the user. The needs of the user are derived from the service level agreement (SLA), decomposed into lower level metrics that can be applied to allocate from a micro-grid of IT, power and cooling resources at the cloud infrastructure level. As an example, consider a cloud data center with a power micro-grid using locally sourced wind, sun, biogas from waste, and natural gas. For a given non-critical service, jobs may be queued up for a time-variant supply source such as wind and solar electricity to come online, while jobs that require immediate servicing are executed using biogas derived from common and novel supplies of power (e.g. from manure at a dairy farm[15]). Indeed, a microgrid of locally-sourced power and cooling resources can enable the cloud services to be supplied from a *net zero data center* – a data center that does not draw any power from the utility and runs at lowest cost using appropriate supplies of power.

Additionally, to achieve this dual vision of improved ecosystem sustainability and reduced service delivery cost, the following four points within the supply-demand framework must also be considered:

- Lifecycle design and sizing of cloud infrastructure to ensure that the infrastructure is designed for optimum performance throughout the lifecycle and not just during peak operation;
- Monitoring of business services and the correlation of these services to performance and sustainability metrics and providing these results to end-users to assist purchasing decisions;
- Global workload scheduling according to sustainability and performance policies;
- Integration of IT demand management with resource supply constraints.

With progress in the above areas, unique attributes of the cloud infrastructure, particularly with respect to the sharing of resources, can be exploited to reduce the overall cost of service delivery while providing a more sustainable solution.

## PLATFORM AS A SERVICE: SECURE CELLS FOR THE CLOUD

HP Labs' Cells-as-a-Service research focuses on the requirements for an *enterprise-grade* virtualized cloud computing infrastructure, i.e. a service which provides quality of service guarantees over the *security*, *isolation*, *reliability* and *performance* of the virtualized infrastructures it generates and manages. Given a physical infrastructure of computing nodes, storage devices and interconnecting networks, the fundamental requirement is to provide service providers with the illusion of a unique, secure and performant *multi-tenanted* infrastructure designed to their specific requirements. In addition, there are a set of properties such an infrastructure service must support to be considered enterprise grade. The key properties include privacy and security, quality of service and performance, flexibility, scalability and resilience to failure.

Service providers need a platform that can construct virtual infrastructures that are multi-tiered, flexible, and that provide few restrictions relative to existing physical infrastructures. To provide the degree of flexibility required within these virtual infrastructures, a service running in the cloud must be able to scale up or down dynamically its resource usage to cope with changing workloads. However, service providers will want to limit the scope of this scaling for reasons such as cost, and to do so in a way that cannot be subverted by a run-away service or one infected by a virus. Consequently, secure out-of-band mechanisms must be provided for service providers to implement policies regarding limits on flexing and other changes that will be enforced by the automated infrastructure management service, but within which a service may change and flex as it requires.

Cloud services require security and performance isolation from each other; their data and sensitive information protected from other services, as well as performance guarantees regardless of what other services share an infrastructure. Furthermore, service management and core infrastructure management must be separated so that services cannot interfere with the operation of the infrastructure management. With this separation of infrastructure management and service management, there must be a clear separation of concerns between the parties to decide on policies around failure recovery and quality of service to eliminate unexpected service behavior. Consequently, the infrastructure service needs to provide mechanisms for the service provider to push policies into the infrastructure service, detailing aspects they wish to delegate to the service. Without this, enterprises and governments would not trust cloud infrastructure services.

Cells-as-a-Service is built upon the fundamental concept of a *Cell*. Cells contain the collection of virtual components, specifically virtual machines (VMs), storage volumes and subnets declared as elements in a Cell model. The model also describes how these components are connected to create the desired virtual infrastructure. Each component or connection may be defined with a set of attributes relevant to that component. For example, VM elements include specifications for memory requirements, bus addresses for volume attachments, connection to one of a number of Cell subnets, and behavior in the event of failure. The Cell model is described in an XML document (Figure 6). Changes to the model may be handled by submitting an updated model document or through an API that supports incremental changes to the model. Cells are created on demand for a service provider via a *Portal*, with each cell created initially containing only a Cell Controller. This Cell Controller is responsible for securely interacting with the service provider and also monitoring the status of the virtual infrastructure.

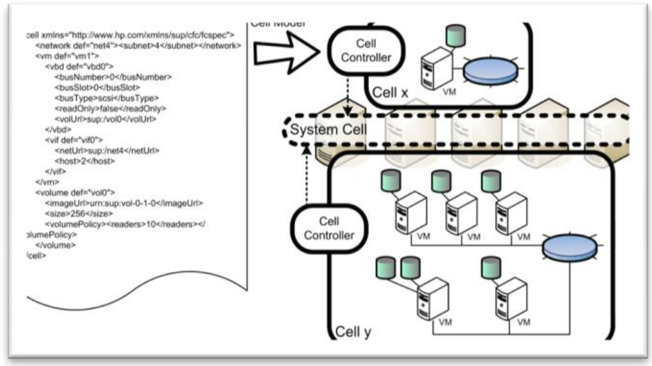

Figure 6: The Cells-as-a-Service Template Driven Architecture

The boundary of a Cell, and indeed any separation of components within the Cell, is secured by the underlying system. This boundary consists of both network connectivity between hosts and subnets of the Cell and between Cells, plus the ability to mount volumes owned by other service providers. By default network traffic is only permitted between VMs on the same subnet and volumes are only visible for connections and imaging within the same Cell. Security may be relaxed in a controlled way by adding rules to the Cell model to share volumes with other Cells and open network connections from a subnet or perhaps only an individual VM to other subnets or VMs in the same Cell. Network connections may also be opened between VMs and subnets in different Cells provided both Cells contain reciprocal rules.

Cell management is entirely handled by a special privileged Cell known as the System Cell. No part of any other Cell may communicate directly with the System Cell apart from a locked-down and secure bastion component of each Cell Controller. The System Cell is responsible for creating and deleting all the virtual components and managing their connectivity, for enforcing all the connectivity policies defined within each Cell and for Cell interaction, and for enforcing any policy regarding recovery or scalability limits associated with each Cell.

The System Cell runs across all the physical hosts, each of which must be running a hypervisor. The System Cell contains two types of component: Host Managers and core system services such as Resource Management and Storage Management. A Host Manager runs on each physical host within the privileged VM (often known as the host OS) and is responsible for managing and validating every action that occurs on that physical host.

Each Host Manager enforces the isolation of Cells from each other and from the System Cell by mediating access to the physical host's compute, network and storage capabilities, and by transforming abstract VM model elements into configuration data appropriate to the underlying hypervisor. The Host Manager interacts with a Storage Manager to create and remove virtual block devices as required by hosted VMs.

Cell subnets are implemented as virtual overlay networks on a single shared physical network. No special hardware is required; instead this has been implemented using a novel fully-distributed virtualized router that facilitates single network hop communication between endpoints; unlike traditional software routers, this solution operates at the OSI network layer so that packets can be forwarded directly to their destination. Every physical host implements a part of the distributed virtual router by filtering and modifying all packets from or to its local VMs according to the network rules expressed in the Cell models. To support the requirement to manage overall performance, the networking layer provides networking resource control to limit and prioritize VM's bandwidth consumption.

## SOFTWARE AS A SERVICE: NOVEL SERVICES FOR THE CLOUD

Web 2.0 and the Cloud have given rise to a new class of services. One of our research thrusts aims to harness the collective intelligence of the connected population inside and outside enterprises to create novel technologies and services. Ours is an interdisciplinary approach, combining sociology, economics, and computer science. We focus on the economics of attention, creating models to understand and harness the flow of collective attention, supporting a mobile society with context-aware and anticipative technology, and enabling a fluid enterprise, where collective intelligence is harnessed via appropriate incentives to predict the future, make decisions, and allocate resources.

The HP *ePrint* platform harnesses the cloud to break down the barriers of distance and connectivity and allow people to send the files they want to print from their mobile devices. From a mom and son printing drawings from an iPad, to an executive on a train sending a presentation from a Palm Pre or BlackBerry® smartphone to print and pick up at a FedEx Office store, HP *ePrint* allows people to securely print anytime, anywhere.

In recent years, social media has become ubiquitous and important for social networking and content sharing. And yet, the content that is generated from these websites remains largely untapped. We have recently demonstrated a service that analyzes the allocation of attention within social media to predict real-world outcomes. We have already tested it by using the chatter from Twitter.com to forecast box-office revenues for movies. The rate at which tweets are created about particular topics can outperform market-based predictors. We further demonstrated how sentiments extracted from Twitter can be further utilized to improve the forecasting power of social media. The methodology is very general and can be applied to any accessible social medium to predict trends in products and services, among others. [16]

We have also developed a method for accurately predicting the long time popularity of online content from early measurements of user's access. Using two content sharing portals, YouTube and Digg, we showed that by measuring the initial rate at which users view and vote on content offered by these services we can predict the long-term popularity of submissions. In the case of Digg, measuring access to given stories during the first two hours allowed us to forecast their popularity 30 days ahead with remarkable accuracy, while downloads of Youtube videos needed to be followed for 10 days to attain the same performance. This methodology can be used on any social medium or web sites that allows for the measurement of download rates. [17]

Advances in Information Technology during the past 20 years have made cloud computing a reality. But many challenging problems remain. These problems include securing services, data and the infrastructure from attack, as well as ensuring that personal data remains private. Flexible and dynamic resource allocation must occur based upon events or policy and at a never before imagined scale. In large complex systems such as these failure is a common attribute and thus services must remain available to clients regardless of hardware or software failure or disruption. Performance must be predictable across a wide range of workload demands while maintaining acceptable economics in delivering the service. New system architectures, programming models, development environments, and testing and debugging methodologies will be required for dynamically instantiated, ephemeral services. The next generation of scientists and engineers will focus on creating and delivering these advances.

## Acknowledgement

This paper provides a broad overview of the work in Cloud Computing being carried out by more than 100 researchers at HP Labs over the past decade in Cloud and Security Lab, Intelligent Infrastructure Lab, Sustainable Ecosystems Group and Social Computing group. Owing to lack of space, we could not list all of them as co-authors. For a listing of the researchers in the different labs and groups at HP Labs, please visit the HP Labs web site at www.hpl.hp.com .

## References

1. HP Labs 2009 Annual Report - http://www.hpl.hp.com/news/2010/jan-mar/annual_report.html
2. "Amazon explains its S3 outage", http://www.zdnet.com/blog/btl/amazon-explains-its-s3-outage/8010
3. S. Gilbert and N. Lynch. Brewer's conjecture and the feasibility of consistent, available, partition-tolerant Web services. ACM SIGACT News, 33(2):51-59, June 2002
4. A. Aiyer, E. Anderson, X. Li, M. Shah and J.J. Wylie, Consistability: Describing usually consistent systems. In HotDep 08, December 2008.
5. E. Anderson, X. Li, A. Merchant, M. Shah, K. Smathers, J. Tucek, M. Uysal and J.J. Wylie. Efficient eventual consistency in Pahoehoe, an erasure-coded key-blob archive. International Conference on Dependable Systems & Networks, June 28-July 1, 2010.
6. E. Anderson, X. Li, M. Shah, J. Tucek, and J.J. Wylie. What consistency does your Key-Value Store actually Provide? Sixth Workshop on Hot Topics in System Dependability (HotDep 2010), October 3, 2010, Vancouver, B.C. Canada
7. Kevin Lim, Parthasarathy Ranganathan, Jichuan Chang, Chandrakant Patel, Trevor Mudge, Steve Reinhardt. Server Designs for Warehouse-Computing Environments. IEEE Micro 29(1), Special issue on Top Picks in Computer Architecture, 2009.
8. Ramya Raghavendra, Parthasarathy Ranganathan, Vanish Talwar, Zhikui Wang, Xiaoyun Zhu. No "Power" Struggles: Coordinated Multi-level Power Management for the Data Center. In Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), Seattle, WA, March 2008.
9. Emerging technologies and their impact on system design, N. Jouppi and Y. Xie, Tutorial at the International Symposium on Low Power Electronics and Design, 2009
10. Parthasarathy Ranganathan, From microprocessor to Nanostores: Rethinking System Building Blocks for the Data-Centric Era, IEEE Computer, January 2011, To appear.
11. Parthasarathy Ranganathan, From microprocessor to Nanostores: Rethinking System Building Blocks for the Data-Centric Era, IEEE Computer, January 2011, To appear.
12. Chopra, A, $35 Computer Taps India's Huge Low-Income Market, Christian Science Monitor, 2010
13. Patel, C., Shah, A, Cost Model for Planning, Development and Operation of a Data Center,,HP Laboratories Technical Report, HPL-2005-107R1, Palo Alto, CA, 2005

14. Chandrakant Patel, Sustainable Ecosystems: Enabled by Supply and Demand Management, International Conference on Distributed Computing and Networking (ICDCN), Jan 2011, To appear
15. Sharma, R., Christian, T., Arlitt, M., Bash, C., Patel, C, Design of Farm Waste-driven Supply Side Infrastructure for Data Centers, ASME-Energy Sustainability, 2010
16. Asur, S., and Huberman, B., Predicting the Future with Social Media, Proceedings of the ACM International Conference on Web Intelligence, 2010
17. Szabo, G., and Huberman, B., Predicting the popularity of online content, Communications of the ACM, August 2010, Vol 53 No 8.