IT@Intel White Paper

Intel IT
IT Best Practices
Private Cloud and Cloud Architecture
December 2011

Best Practices for Building an Enterprise Private Cloud

As we begin the final phases of development for our enterprise private cloud, we have identified best practices in several areas that have helped us maximize the business benefits of cloud computing, introduce solutions quickly into our environment, and pursue our business goals.

Sameer Adhikari
Cloud Business Intelligence Architect, Intel IT

Greg Bunce
Automation Engineering Lead, Intel IT

Winson Chan
Software Engineer, Intel IT

Ajay Chandramouly
Industry Engagement Manager, Intel IT

Das Kamhout
IT Cloud Lead, Intel IT

Brian McGeough
Managed Cloud Engineering Manager, Intel IT

Jon Slusser, Jr.
Cloud Services Operations Manager, Intel IT

Catherine Spence
Enterprise Architect, Intel IT

Bill Sunderland
IT Cloud Foundation Engineering, Intel IT

Executive Overview

Intel IT has achieved significant progress on our multi-year initiative to build an enterprise private cloud. By implementing a cloud strategy, we have saved USD 9 million to date—and we anticipate approximately USD 14 million net present value (NPV) over the next four years.

Most Intel business groups now run production applications in our private cloud, giving them capabilities they did not have in the past. Through extensive automation, we have made self-service infrastructure provisioning the norm, delivering 80 percent of new servers inside our cloud in less than three hours, and most within 45 minutes.

We are actively expanding our use models for cloud beyond infrastructure as a service (IaaS) and have outlined a roadmap to implement hybrid clouds for even higher agility and efficiency. As we begin the final phases of development for our enterprise private cloud, we have identified best practices in several areas that have helped us maximize the business benefits of cloud computing, introduce solutions quickly into our environment, and pursue our business goals. These goals include:

- 80-percent effective utilization of IT assets
- Consistent increase in business velocity
- Zero business impact from IT infrastructure failures

Going forward, these best practices will continue to unlock the full power of cloud computing at Intel.

Contents

IT@INTEL

The IT@Intel program connects IT professionals around the world with their peers inside our organization – sharing lessons learned, methods and strategies. Our goal is simple: Share Intel IT best practices that create business value and make IT a competitive advantage. Visit us today at www.intel.com/IT or contact your local Intel representative if you'd like to learn more.

BACKGROUND

As part of our overall data center strategy, Intel IT is creating business efficiencies by vertically integrating and optimizing our data center infrastructure to align with the different requirements of Intel's critical business functions: Design, Office, Manufacturing, Enterprise, and Services (DOMES). As shown in Figure 1, cloud computing is a key initiative for Office, Enterprise, and Services.

We are well into a multi-year project to implement an enterprise private cloud that supports our Office, Enterprise, and Services environment applications, such as access to databases, files, and the Web. We have standardized our cloud infrastructure on a strategic mix of Intel® Xeon® processor 5500 series-based servers and Intel® Xeon® processor 5600 series-based servers. Although our cloud is not yet complete, we have achieved significant accomplishments so far, most notably in the area of supplying compute infrastructure through self-service provisioning.

As we move into 2012, we are shifting our focus to deliver data and application platform services through the cloud. Our goal is to develop an enterprise private cloud that delivers a highly available computing environment—providing highly secure services and data on-demand to authenticated users and devices from a shared, elastic, and multitenant infrastructure. Beyond that, our strategic vision includes a robust hybrid cloud environment that enables our strategic business goals:

- **80-percent effective utilization of our IT assets.** We intend to increase the utilization of resources and lower our capital and operating expenses through a combination of pervasive virtualization—including secure virtualization of enterprise applications—and increased sharing among tenants through larger pools of VMs in fewer data centers. We foresee the eventual ability to share all Intel resources across all business units to enable higher capacity reuse. We will accomplish this by creating large fungible pools of resources with specialized hardware where appropriate.
- **Consistent increase in business velocity.** We want on-demand self-service to become the norm within the enterprise. In the future, we anticipate expanding this concept by provisioning virtual machines (VMs) within minutes and providing complete application environments with one button or API call. We also envision using the external cloud to handle peaks in demand as well as automating sourcing decisions. All of these efforts support a larger goal of enabling developers to turn innovative ideas into production services in less than a day.
- **Zero business impact from IT infrastructure failures.** Using a combination of cloud-aware applications designed for failure and an automated end-to-end service managed cloud, we can help

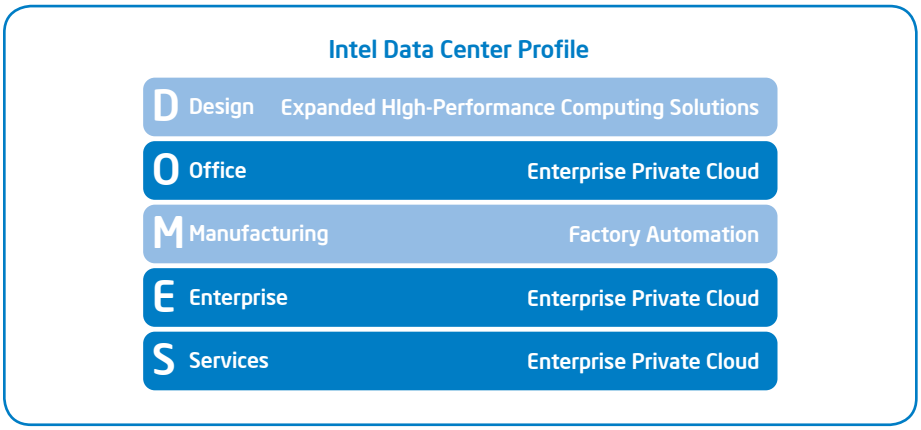

Figure 1. Cloud computing offers significant benefits for the Office, Enterprise, and Services business functions.

ensure that our services are always running, even when we have infrastructure failures.

Our private cloud offers key benefits to Intel:

- **Agility.** We deliver 80 percent of new servers in less than three hours, and most within 45 minutes. In contrast, just two years ago, server provisioning in the traditional IT environment typically took as long as 90 days.
- **Business intelligence.** We measure system utilization and health based on actual data, and we provide transparent data to Intel's business groups to help them with resource planning.
- **Elasticity.** We are in the final phase of enabling dynamic scaling of resources based on user demand, making our infrastructure more efficient and responsive to customer needs. Users can currently scale up using our self-provisioning portal, and over the next six months we anticipate deploying more automated elasticity, enabling scale-out and scale-back to achieve peak capacity on demand.
- **Cost efficiency.** Through multitenancy and resource pooling, the private cloud supports higher effective utilization of resources. Our cloud investments have already paid for themselves and have returned USD 9 million in cash savings (not including non-cash business value such as employee productivity). Before our move to the cloud, our average utilization rate was about 8 percent. Cloud computing has increased that to 40 percent, and we anticipate further improvements in effective utilization as we complete our enterprise private cloud.

Additional benefits include a significant reduction in infrastructure footprint—including servers, racks, network ports, and I/O ports—as well improved network manageability, improved disaster recovery abilities, and reduced greenhouse gas emissions and energy consumption.

Overview of Intel's Cloud Challenges and Solutions

During the last two years, the cloud computing industry has matured significantly. Many off-the-shelf options are available now that did not exist when we began our private cloud initiative. We therefore developed specific applications that met our cloud computing needs and allowed us to reuse the existing investments we had made in managing our enterprise IT environment.

We faced challenges in a number of areas as we developed Intel's enterprise private cloud:

- Integrated cloud foundation
- Pervasive virtualization
- End-to-end system health monitoring
- Elastic capacity and measured services
- On-demand self-service

We continue to develop solutions to some of these challenges as we enter the final phase of private cloud implementation and deployment.

INTEGRATED CLOUD FOUNDATION

Similar to managing a supply chain, we integrated data from our cloud foundation—network, storage, compute, applications, and facilities to create a compute infrastructure as a service (IaaS). We eliminated data silos to create a single data warehouse, consumable by both employees and automation tools, to enable real-time decision making. We developed an in-house solution that provides a highly reliable and cost-effective foundation for our cloud, utilizing traditional database technologies.

As shown in Figure 2, we can use the integrated data to predict problems before they lead to an outage or quickly fix them when they arise.

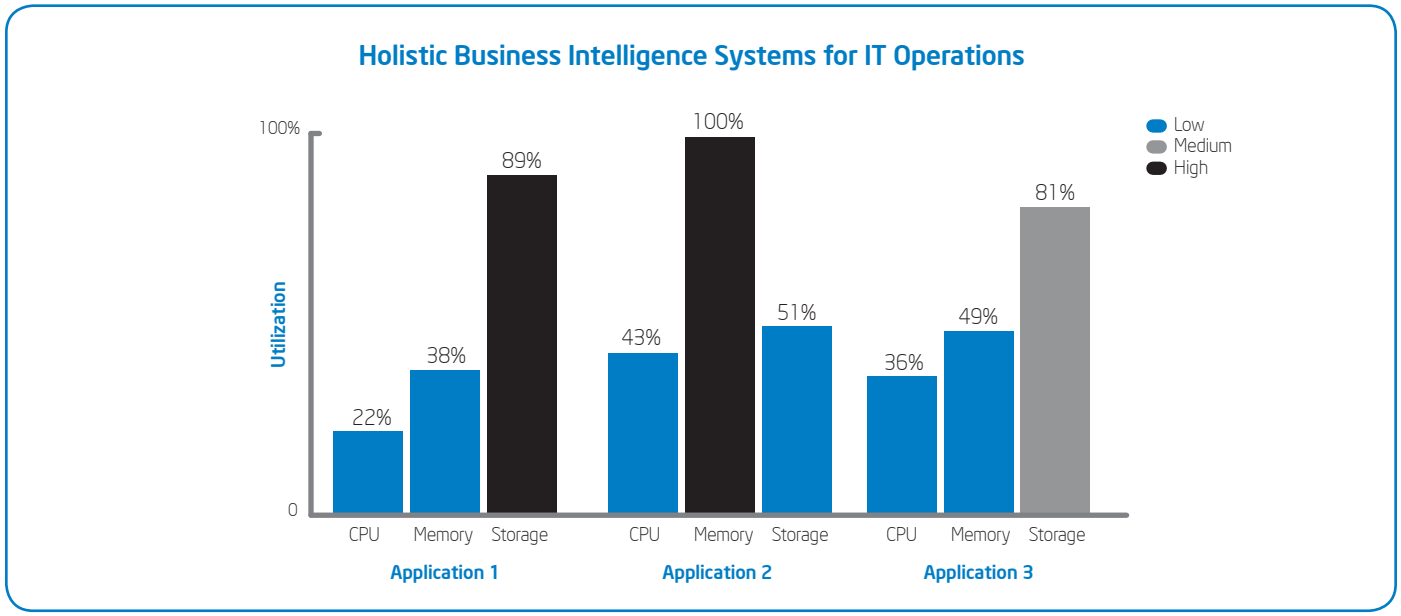

Figure 2. Our integrated data information can alert us to CPU, memory, and storage utilization issues on a per-application basis.

PERVASIVE VIRTUALIZATION

When we decided to build our private cloud, we had already begun implementing virtualization. However, our efforts were on a relatively small scale: Virtualization was the exception rather than the rule and was typically undertaken in response to specific requests from business groups. As a result, by the end of 2009, we had virtualized only about 12 percent of the Office and Enterprise environment. Our capacities and capabilities were constrained due to technical limitations; we were able to virtualize only smaller and less-demanding systems that comprised less than 50 percent of the environment.

Also due to technical limitations, some early virtualization experiences were disappointing—making IT management reluctant to support full-scale virtualization. Another reason for the slow progress was that, although our architects and engineers had designed virtualization solutions, our operations teams were not actively engaged with business groups to explain our plans and encourage them to virtualize.

Three factors helped create momentum for our virtualization efforts:

- IT management set a clear mandate to make virtualization a priority.
- Expanding technical capabilities enabled us to expand the scope of our virtualization efforts while maintaining quality of service with no negative impact to the production environment.
- We proactively demonstrated that virtualization delivered significant business value.

END-TO-END SYSTEM HEALTH MONITORING

Intel has used external suppliers to host some parts of Intel applications, but monitoring these externally hosted applications has historically been of interest only to our application development teams. As our enterprise private cloud matures, however, and we begin to provide services in the cloud, we must be able to monitor these externally hosted applications in order to measure overall service availability.

End-to-end cloud monitoring goes beyond the normal complexity of monitoring distributed systems because we have to integrate internal and external monitoring of systems and services. We must also define new roles for operations support staff to transform our support service-level agreements (SLAs) into true end-to-end monitoring. In addition, because monitoring encompasses many areas of expertise and support organizations, including facilities, network, storage, compute, and applications, we are establishing a sense of joint ownership across our IT Operations teams and internal customer software development teams.

To help address these issues, we are developing a tool that helps to define service-health models and pulls the monitoring data from each of the component-level domains. The tool then dynamically provides a service-level dashboard that enables an IT Operations support team to see if a service is operating within normal, failed, or degraded conditions. The support team can use the dashboard to monitor the transitions in and out of such states. By combining reports from several monitoring systems, we can obtain a thorough view of the user experience and event severity.

We have introduced the new monitoring tool to a limited number of application teams through a series of proofs of concept (PoCs). The PoCs have illustrated that this new capability delivers a cleaner incident management ticketing process and greatly reduces noise alerting and manual error.

By providing an automated, one-stop-shop for incident management ticketing, the new tool relieves IT Operations support teams of having to scroll through thousands of records of irrelevant event data and deciding what to ticket. This frees the teams to focus on the more productive and demanding task of managing the increasing complexity of application and service deployments both internally and externally, with no incremental investment

ELASTIC CAPACITY AND MEASURED SERVICES

The ability to scale quickly based on demand requires the ability to detect demand increase, understand what capacity is available for scaling out, and view this data historically to determine utilization norms and peaks. Although users of public clouds generally perceive infinite scalability, in reality, nothing in a data center—or even a group of data centers—is infinite. A goal for our private cloud is to create a similar illusion of infinite ability to scale to meet business requirements.

We have developed methods to determine resource requirements for business group applications, quickly provide those resources, and establish a way to measure services so we can report usage information back to the business groups over time. As in other areas of the overall cloud solution stack, this involves deploying both automation and business intelligence tools to allow for reactive elasticity and retroactive demand analysis to help us understand the nature of each application's demand.

Beyond elasticity at this fundamental level, we need to provide tenants in our private cloud with the best quality of service possible. Consolidation of multiple VMs onto a single host creates a unique performance/capacity management problem that many traditional manageability solutions do not adequately address. The risk of oversubscribing VMs and saturating the underlying I/O, memory, or CPU can lead to degradation in the quality of service. We have addressed this new complexity by enabling transparency into the aggregation of subsystem usage and mapping those views to the applications or services running on the platforms and/or infrastructure.

As an example of how our integrated data provides improves decision making, we detected that at 4:00 p.m. every day many of our VMs were having I/O latency issues. By correlating and analyzing the integrated data, we pinpointed a process

running on many systems simultaneously, which caused an I/O storm. Before data integration, we would have seen this data as a silo and would not have been able to understand the magnitude of the issue. Nor would we have had the ability to correlate the VM performance to the I/O constraints.

ON-DEMAND SELF-SERVICE

To be effective, a cloud needs to be automated so that customers can provision resources themselves, eliminating manual processes. In addition to the back-end automation work required to create such a system, we also developed a simple, intuitive, and, most importantly, customizable front end. In particular, we did not want users to have to interact with several portals to self-provision different parts of the infrastructure, such as storage, VMs, and network resources. Instead, we wanted to create a "one-stop shop" for all infrastructure services.

Workflow automation is a major aspect of implementing virtualization across the enterprise. We adapted business processes that take advantage of automation tools in the environment. We also identified gaps where automated tools weren't sufficient to resolve issues when they arose. We realized that the more we relied on automation, the more important it was to have consistent processes and infrastructure.

While our Design and Manufacturing environments are very automated at scale, our Enterprise IT shop has traditionally performed many tasks manually. We replaced a number of manual steps and functions the IT operations group had performed—a significant undertaking. In addition, we integrated business, technical, and user-related processes within the same automation framework.

Because no off-the-shelf workflow automation solution met all our needs, we built our own workflow automation framework, which can access standard interfaces as necessary to support other cloud solution layers such as health monitoring and on-demand self-service provisioning.

BUILDING A PRIVATE CLOUD STRATEGY

Building Intel's enterprise private cloud is a complex, multi-year process that requires a comprehensive strategy. We decided to build our cloud from the inside out, focusing first on implementing private IaaS. Now we are turning our attention to providing platform as a service (PaaS) and software as a service (SaaS) for specific use cases.

Our cloud strategy is driving considerable cost efficiencies. By achieving average server consolidation ratios of up to 20:1, we have saved USD 9 million to date—and we anticipate approximately USD 14 million net present value (NPV) over the next four years.

Implement Compute IaaS First to Enable Enterprise Usage Models

Cloud computing creates a pool of compute resources located across multiple sites. This lets us apply Intel's global computing resources to individual projects and provide more compute capacity by increasing utilization of existing resources while reducing the need to add hardware. We are creating flexible pools of compute resources based on newer, more powerful, and energy-efficient servers. Virtualized workloads can be dynamically allocated and migrated between physical servers within these resource pools, and we can divide these pools logically among Intel core business verticals, according to business need.

Deploy PaaS to Speed Application Development

PaaS enables software engineers to build and host custom applications in the cloud. The developer controls the application and hosting configurations, while the underlying infrastructure is abstracted. Platforms contain standardized software, middleware, business Web services, development tools, and automation, which promote a faster path to production and time to market.

The PaaS environment itself is self-service, on-demand, highly elastic, and multi-tenant, with broad accessibility and resource pooling. Further, PaaS facilitates the creation of applications that incorporate cloud characteristics such as elasticity, multi-tenancy, and design for failure. Software developers use a cloud-based environment to create custom, cloud-ready applications.

At Intel, PaaS is being deployed on IaaS in the enterprise private cloud. PaaS will build on and extend the capabilities offered in IaaS. We are focusing on two application stacks for traditional and open-source programming languages. We expect to provide automated elasticity at all layers of an application written to PaaS and build applications and platforms designed for resiliency and no perceivable downtime. PaaS will allow us to increase utilization of the computing environment by compartmentalizing code as opposed to using server virtualization exclusively. Our ultimate goal is to enable business agility to turn innovative ideas into production services in less than a day through further automation and investment in PaaS.

Pursue Federated Clouds to Maximize Agility and Efficiency

Cloud federation is key to maximizing agility and efficiency, enabling easy movement of workloads and services between different IT infrastructures within the enterprise—and ultimately moving workloads between private and public clouds. To reap the benefits of federation without undue IT overhead, applications and service providers need to implement relevant standards for workloads, APIs, tools, and security. Enhancing security to protect data and intellectual property is a key aspect of enabling federation. Intel is helping facilitate the development of the necessary enterprise requirements through the Open Data Center Alliance.

We have laid the foundation to use hybrid clouds to further increase scalability and provide burst capacity. We have been sharing capacity across multiple resource pools; in 2012 we plan to share capacity across data centers and then expand to hybrid use of secure external clouds.

BUILDING A COMPREHENSIVE MANAGED CLOUD CAPABILITY

Before we could begin to integrate the cloud foundation, we needed to understand the types of data associated with each major resource area. For example, we needed to know how much capacity customers were using, the rate at which demand was growing, and if performance was meeting expectations.

Once we categorized the data, we applied Information Technology Infrastructure Library* (ITIL) standards to manage our cloud operations at each level of the cloud capability stack, as illustrated in Figure 3. These standards provide a holistic view of our cloud—including physical devices and containers, functions management, data integration and analytics, workflow automation, and service delivery. Prior to using ITIL standards, we could only view data in silos, which limited our understanding of our IT assets and how to best utilize them.

Functions Management Layer

The functions management layer serves two main purposes—to watch and to act—in four main areas: configuration management, event management, change management, and capacity management. Besides providing a view across the compute environment, the functions management layer also includes a list of actions that can be applied to a device or a container. For example, actions could include provisioning a new OS in a VM, updating and patching an already provisioned OS in a VM, or updating or installing new applications in a VM.

Data Integration and Analytics Layer

The data integration layer allows us to integrate data from all containers as well as from the functions management layer. The data integration layer also feeds into the workflow automation and service delivery layers where appropriate.

Figure 4 shows how we aggregate capacity and performance data into our Infrastructure Management Operational Data Store (iMODS)

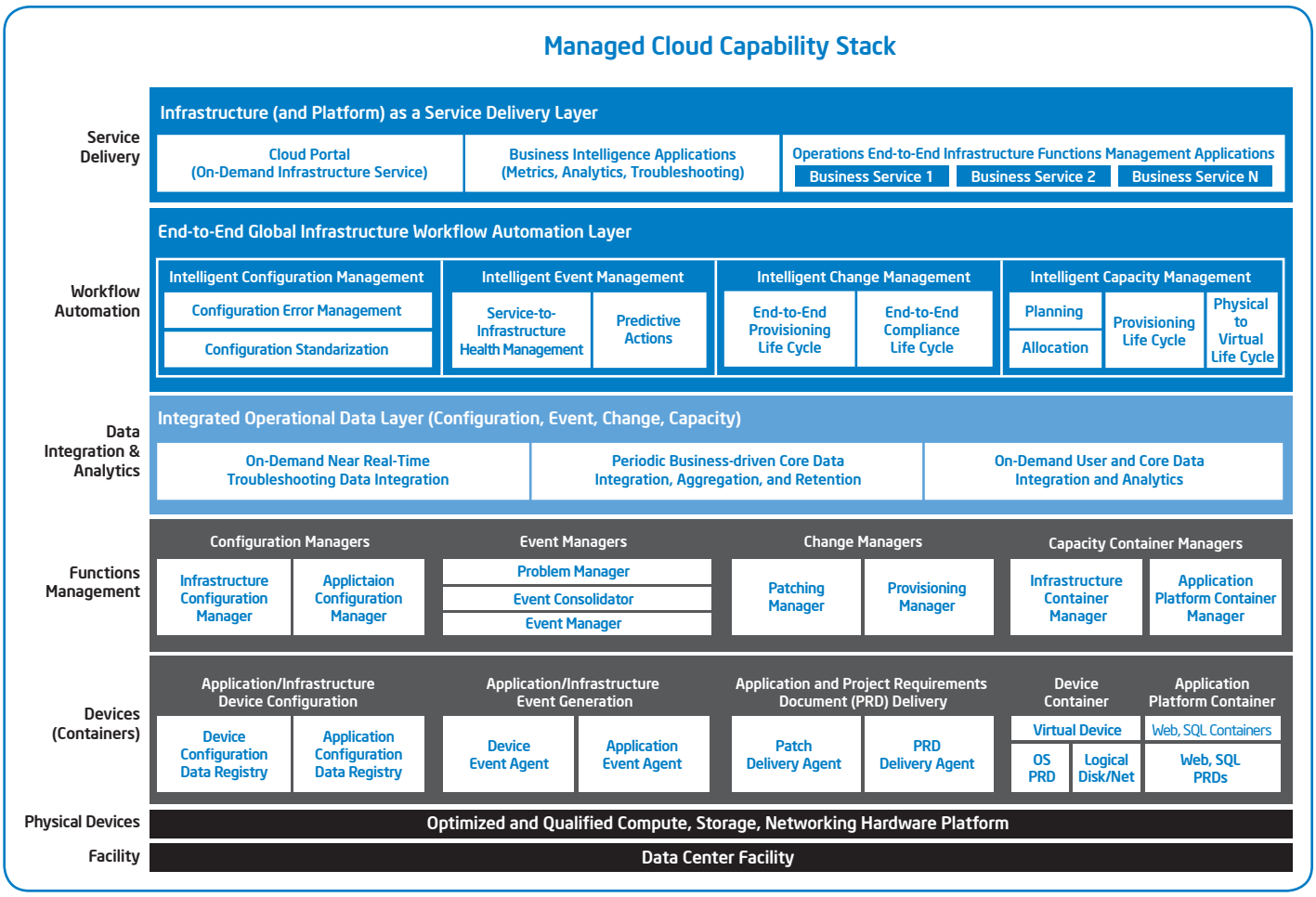

Figure 3. Information Technology Infrastructure Library* (ITIL) standards provide a holistic view of our entire cloud solution stack.

from the different function managers. This data is used by other components in the cloud framework, as shown in the top two boxes in Figure 4. The goal is to unify the data to present a comprehensive picture of cloud performance.

Each function manager is a silo that is ignorant of the others—so unifying the data presents challenges at the extract, transformation, and load levels.

We solved some significant data integration challenges while developing the integration layer.

- **Integrated capacity and performance data.** We needed to integrate capacity and performance data sets for a complete picture of operational capacity. For example, there may be space available on the storage area network (SAN), but the I/O metrics may signal a limit on using the space.
- **Unique entity identifiers.** Although a user may consider "C:" a unique name for a drive on a server, the storage manager or the VM manager does not use this same identifier. Also, a server name can be reused for different machines over time. Data collected for a particular name, therefore, is not always associated with the same device.
- **Consistent end-to-end view of data.** We had to integrate data from different systems. For example, to follow a chain from VM to File Systems to VM Disk Files to Virtual Storage Container to Physical Logical Unit Number (LUN) to SAN Frame, we need to align both performance and capacity data along all links in the chain.

Workflow Automation Layer

An on-demand, highly available, and scalable cloud computing infrastructure requires an equally robust automation capability to match the level of expected agility, elasticity, and efficiency for rapid VM provisioning and de-provisioning. For our private cloud automation, we designed a modular, extensible framework that simplifies integration of the many aspects of cloud computing, summarized in Table 1.

Table 1. Workflow Automation Considerations

Complex business logic	▪ State management, including the ability to resume from an inactive state to accommodate system failure
Ease of design and deployment of workflows	▪ Progress tracking
Availability of tools for systems integration	▪ Post-back of run-time generated data
Management of workflow instances	▪ Support of long-running transactions
Asynchronous programming design patterns, with the ability to handle events and callbacks	▪ Tremendous amount of integration between heterogeneous and disparate systems, including configuration management database, virtual machine management, network management, monitoring, and patching

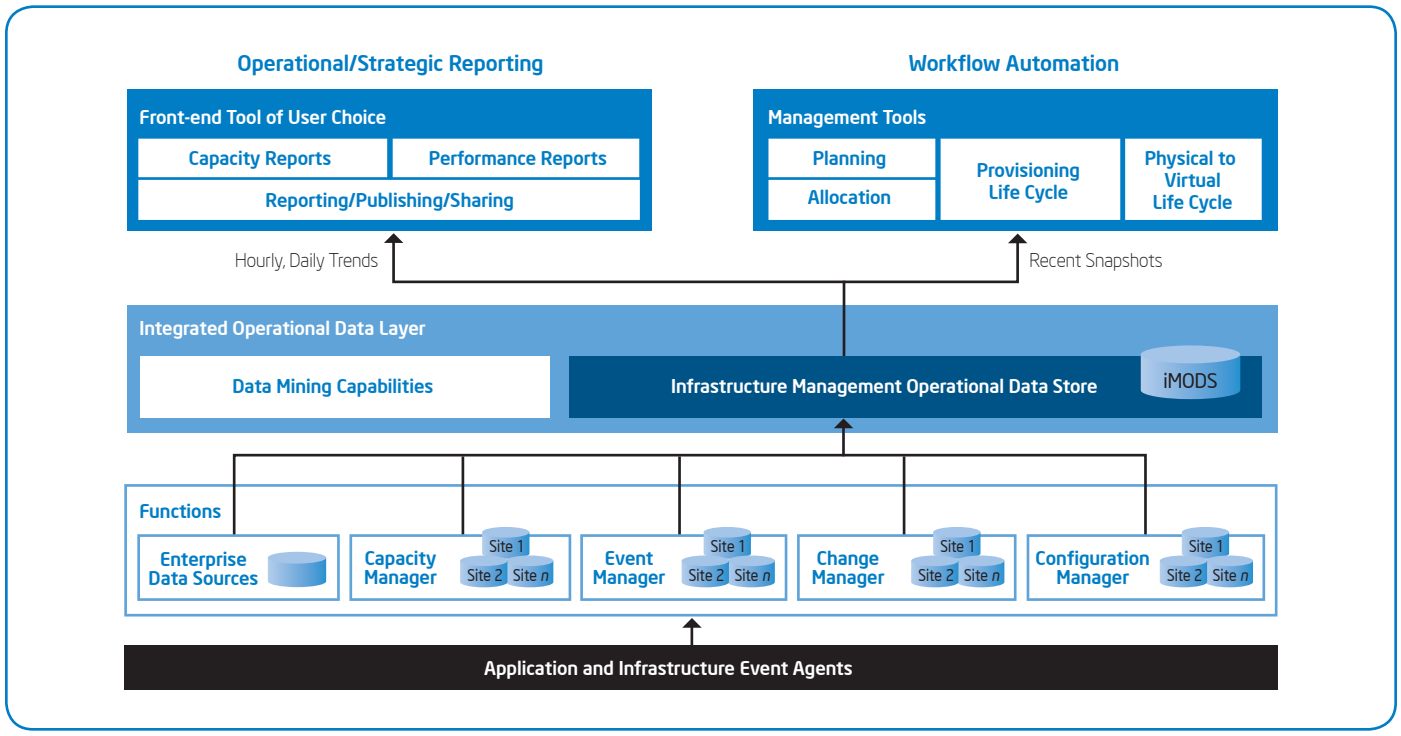

Figure 4. The Infrastructure Management Operational Data Store (iMODS) unifies the data from all the function managers so it is available to end users and management tools.

As the industry matures, the modular nature of the automation framework will allow us to replace components with solutions that match our business and technical requirements. When developing our workflow automation framework, we streamlined and standardized existing processes by examining each step in the manual process. We identified steps where we can reliably repeat with standardization, decisively removed steps that were no longer necessary, explored different approaches to processes, and constantly pushed the boundary of what we can automate.

Our workflow automation framework results in self-provisioned VMs that are fully functional, with compute, network, and storage resources configured to specification. The provisioning includes configuration of machine identity, access controls, and manageability features such as patching, monitoring, and backup—all prerequisites to provisioning VMs in a managed enterprise environment.

The knowledge and experience that our development team gained during early rounds of automation paved the way for more comprehensive automation that includes areas such as release management, load balancers, firewalls, and more complex VM collections.

Service Delivery Layer

At the top of the cloud foundation layer of our private cloud, we expose the entire solution stack through service delivery portal solutions that rely on the data feed from below. The service delivery layer controls the actions that can be taken by different consumers, based on their roles. For example, an engineer can request compute, storage, and networking resources from the cloud portal, whereas an IT Operations manager can access a variety of business intelligence tools to measure actual usage against requested resources or troubleshoot problems.

IMPLEMENTING PERVASIVE VIRTUALIZATION

For many application workloads running in a very heterogeneous enterprise, virtualization is the first logical and practical step to building a cloud. Virtualization supplies the foundation for multitenancy and resource pooling, thereby enabling the sharing of compute and memory resources, and increasing effective utilization.

In 2010, we more than tripled the number of virtualized servers in our Office and Enterprise environment to 42 percent. In 2011, we have virtualized about 60 percent, and our goal is to virtualize 75 percent. To achieve this, we are developing solutions to several technical obstacles that have prevented virtualization of several categories of applications.

Establish a Standardized, Repeatable Process

To accelerate the pace of virtualization, we focused on creating a standardized, repeatable process—analogous to a factory production line—that would enable our IT Operations group to create demand among business groups and quickly virtualize a large number of servers in a highly efficient, predictable way. We call this the Virtualization Factory.[1]

The Intel IT Virtualization Factory consists of seven clearly defined steps that encompass the entire virtualization process:

- **Identification.** Identify servers that are virtualization candidates and confirm their suitability by communicating with server and application owners.
- **Create demand.** Create awareness and demand for virtualization among business groups.
- **Scheduling and capacity management.** Schedule a conversion time and reserve hardware capacity.
- **Conversion.** Perform physical to virtual conversion.
- **Test and debug.** Assist customers with post-virtualization testing and tuning.
- **Standardize configuration management.** Accurately transfer all administrative and systems management functions to the virtualized server.
- **End of life.** Remove physical servers from the environment.

Systematically Determine and Resolve Technical and Business Limiters

To achieve our goal of virtualizing 75 percent of applications, we needed to remove technical limitations so that we could virtualize mission-critical and externally facing applications in a secure manner.

In the Identification step, we discovered applications that could not be virtualized because we lacked a technical solution. We gave the details to the Intel IT Engineering group, which then developed the appropriate solution.

In 2010, we removed technical barriers to virtualization for the majority of Office and Enterprise servers. We enabled mission-critical features such as clustering, database mirroring, and Web load balancers, large VMs with greater than 16 GB of RAM, and virtualization of externally facing application servers.

In 2011, we focused on removing even more technical limiters such as securing applications, storage replication, backup and recovery limitations, very large VMs with greater than 48 GB of RAM, and Sarbanes-Oxley compliance.

To accelerate virtualization, we also needed to create demand among business groups and show them that we could virtualize applications without impact to the production environment. We conducted a broad internal marketing campaign to educate business groups about our plans and the benefits of the private cloud. We also formed close relationships with business groups to accelerate the virtualization process.

Use Private Virtual LANs to Provide Separation Between Applications

To provide a level of segmentation for our virtual environment similar to our physical environment, we needed to establish separation between applications running on a host. Within each virtualization host, we isolate applications by providing a dedicated private virtual LAN (PVLAN) for each application.[2] This provides application network isolation similar to the protection provided within our traditional, non-virtualized computing environment. All communications between applications pass through a firewall or other gateway. This applies even to communications between applications sharing the same host. Network segmentation thus helps prevent compromises from spreading from one VM to others on the same or different hosts.

Deploy Secure Landing Zones

As shown in Figure 5, we deploy virtualization hosts into secure demilitarized zone (DMZ) and secure internal zone (SIZ) landing zones within the virtualized environment. These are comparable to the security zones that exist in our traditional physical Office and Enterprise environment.

Virtualization hosts that support externally facing applications reside in DMZs. Each DMZ landing zone consists of a subnet protected from the Internet by a firewall, and separated from the intranet by a separate set of firewalls.

Virtualization hosts that support secure internal applications reside in SIZs. Each SIZ landing zone consists of a subnet protected from the internal intranet by a set of firewalls. A SIZ landing zone may also host other systems, such as management servers, that communicate with applications in the DMZs.

These security zones help prevent a successful attack from spreading through the virtualized environment. With this model, multiple VMs and landing zones would have to be breached to result in widespread compromise of the environment.

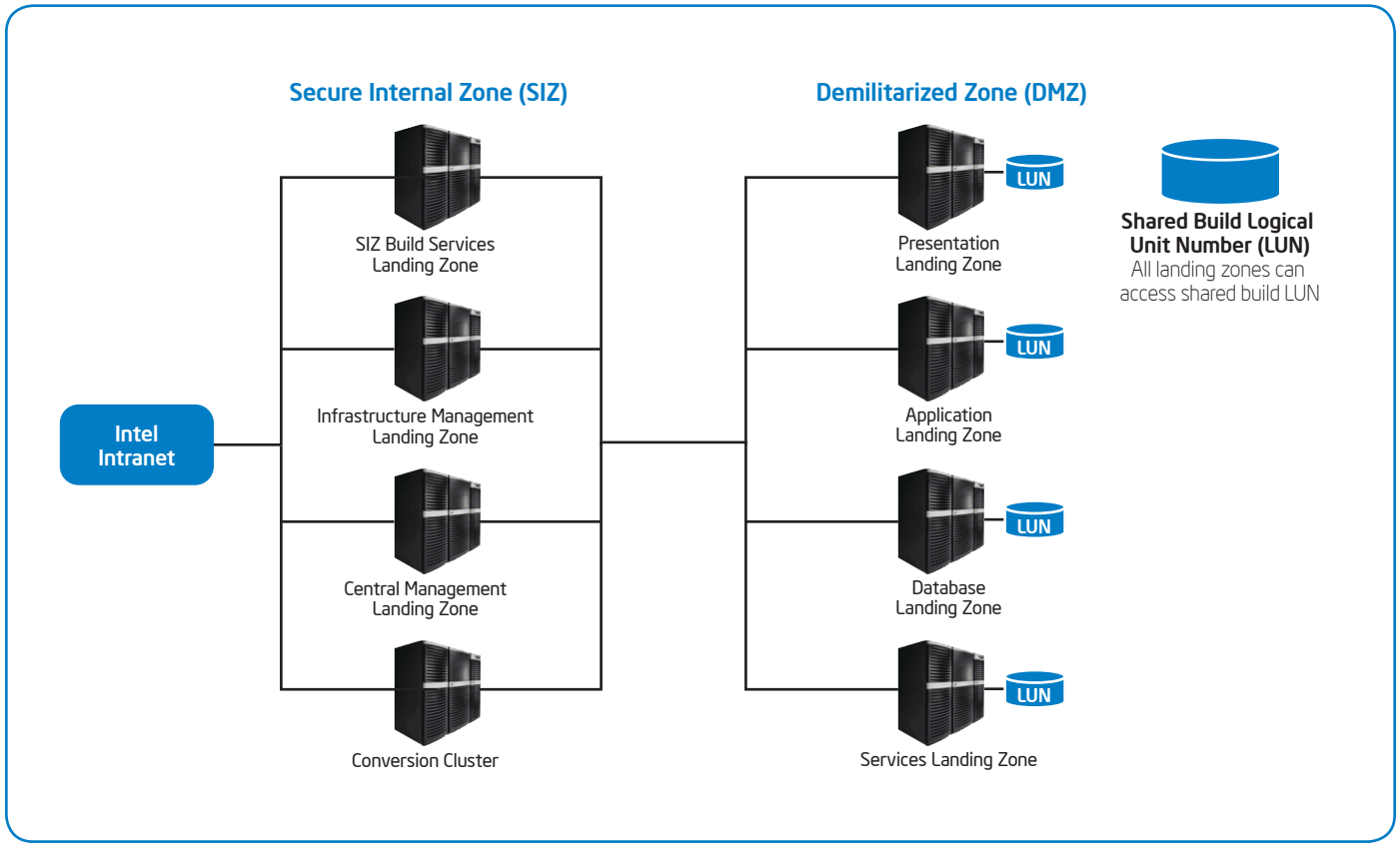

Figure 5. Demilitarized zones (DMZs) and secure internal zones (SIZs) help prevent a successful attack from spreading through the virtualized environment.

ESTABLISHING END-TO-END HEALTH MONITORING

Within Intel IT, the growing complexity and volume of services we needed to support was starting to look similar to other areas of our business that scaled quickly. Considering this growth, we wanted to be better able to monitor the health of services, both inside and outside the firewall, without increasing headcount to meet the demand. We consider end-to-end health monitoring to be one of the most critical elements of managing an enterprise cloud environment.

End-to-end health monitoring provides the ability to quickly respond to complex component and application issues when they arise—resulting in increased user satisfaction and cost savings. It is usually less expensive to fix a problem early, before it becomes complicated or widespread. End-to-end health monitoring also results in a more reliable infrastructure, which can translate to greater productivity.

By applying the following best practices, we have improved our alert-to-ticket ratio by 50 percent and reduced IT Operations support headcount by 100 hours per week. As we continue to enhance our workflow automation and advanced analytical capabilities, we anticipate even further operational efficiencies.

Provide Manageability through an Automated Virtualized Infrastructure

Adding manageability through an automated virtualized infrastructure is the foundation for enabling the health monitoring service to respond dynamically to changing business direction. We have built an automation framework that is capable of managing all facets of manageability throughout the life of a service:

- **Provisioning.** We automate rapid deployment of new OS instances, monitor load for placement management, and integrate the responses of multiple provisioning tools.
- **Distribution of software and configurations.** At this layer, we package software for rapid deployment, manage repair and reconfiguration responses, and configure manageability as we deploy service components.
- **Data and ticket events.** At this layer, we integrate multiple event correlation capabilities, include event-to-service model correlation for extrapolation, detect state changes that trigger automated responses, and limit manual responses to events that have entered a service queue.

Use a Phased Approach to Implement Monitoring

Implementing end-to-end health monitoring is a complex task that affects many IT processes. Therefore, we adopted a phased approach, with clear goals, milestones, and success metrics for each phase. We started with only a few key IT managed services, such as our externally facing business-to-consumer services, instead of trying to monitor all services from the very beginning.

- **Phase 1: Event Noise Reduction and Health Monitoring.** In this phase, we concentrated on developing an event management process with a clear set of roles and responsibilities, automating ticket creation, developing a service health model and event identification process, and enabling component- and service-level event handling. The model enables us to detect symptoms of a problem early, correctly diagnose the cause, and fix it before the application performs unacceptably.
- **Phase 2: Responsive Incident/Impact Resolution and Intelligent Impact Avoidance and Self-healing.** In this phase, we will focus on the following:
  - Taking effective and largely automated actions to restore services quickly.
  - Providing continuous service quality monitoring and taking actions to deliver high service availability.
  - Performing predictive analytics of potential impacts and incidents.
  - Taking proactive and automated actions to avoid major impacts as well as enable effective reactive healing.

Include All Key Elements

As we continue to build our end-to-end health monitoring system for our private cloud, we consider all of the following:

- **IT service management.** Enable, improve, and maximize the delivery of IT services that support Intel's business.
- **Service impact management.** Monitor and manage the performance and availability of business-critical services, and visualize service relationships
- **Application management.** Analyze how to gain the most benefit from Intel's enterprise resource planning (ERP), customer relationship management (CRM), and other applications.
- **IT operations, database, and infrastructure management.** Build a solid foundation for successful business service availability monitoring, which includes open solutions that interoperate with third-party enterprise management technologies.

PROVIDING ELASTIC CAPACITY AND MEASURED SERVICES

Measured services is a key attribute of cloud computing. To help us reach our goal of 80-percent effective utilization, we set up systems to monitor and optimize resources. We also implemented process improvements to increase automation and to eliminate manual operations that impeded efficiency and agility.

Prior to implementing our private cloud, we experienced several problems associated with capacity:

- Low resource utilization and unbalanced capacity constraints
- Little accountability for efficient use of capital investments
- Largely manual operations that hindered efficiency and agility
- Disruption of services due to a lack of actionable data and reactive resolution of performance issues

Automation of capacity planning and management can improve productivity and time to market (TTM), as well as resource utilization, cost efficiency, and system predictability. In addition, it encourages desired customer behavior through transparent data about allocation, usage, and cost. We use the following definitions:

- **Capacity planning.** Providing IT infrastructure at the right time in the right volume at the right price, and identifying how it can most efficiently used.
- **Capacity management.** Providing a point of focus and management for all capacity and performance issues related to services and resources.

As shown in Figure 6, our capacity management model is circular, starting with forecasting, planning, and budgeting. We then purchase resources based on the outcomes of those steps. We also include a reconciliation step that provides feedback to the forecasters—enabling them to learn from what we've already done and better forecast in the future.

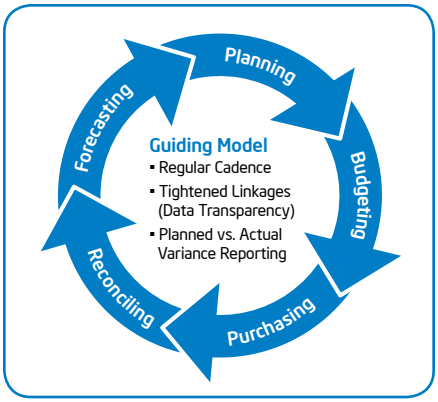

Figure 6. Our model for providing elastic capacity and measured services includes providing data back to the forecasters to enable better decision making in the future.

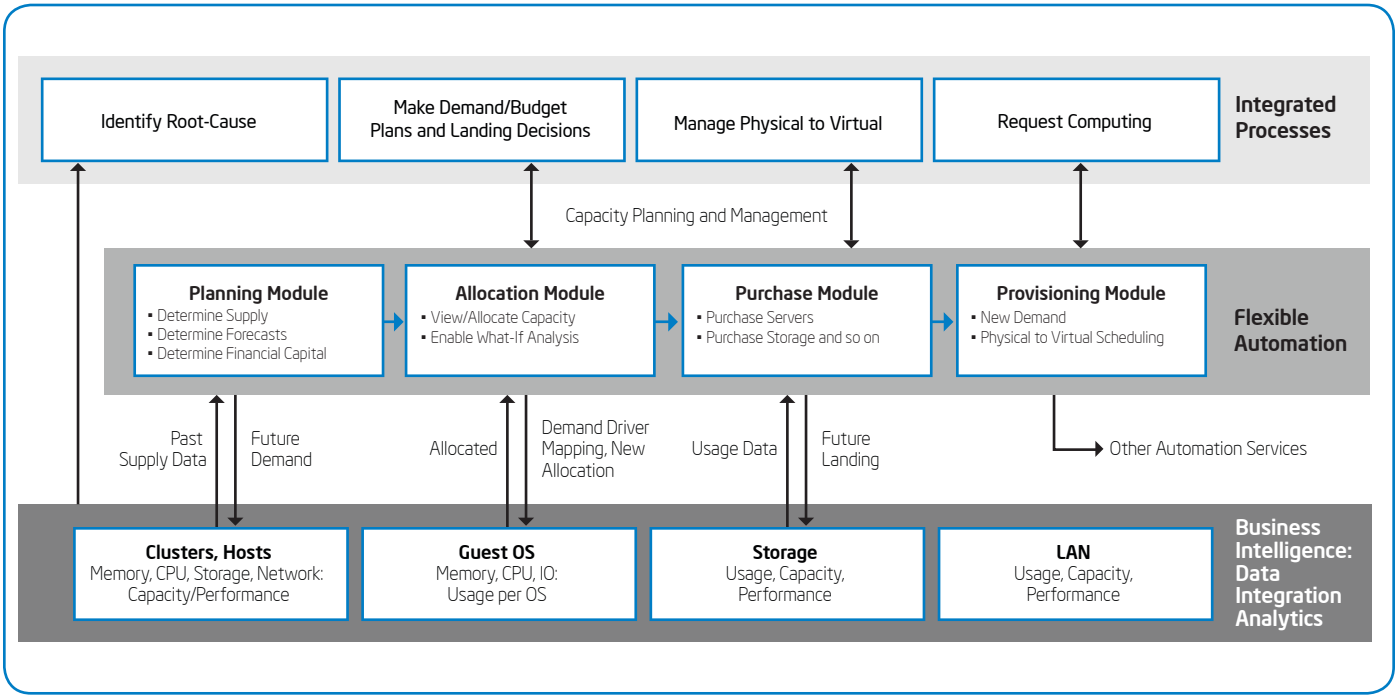

Figure 7. Our capacity management infrastructure uses business intelligence, workflow automation, and integrated processes.

We also have defined two types of use cases:

- "Ad-hoc" demand, in which consumers are looking for a small amount of capacity on-demand with no forecasting.
- Forecasted demand, which allows consumers to utilize much larger chunks of capacity in order to meet larger scale demand needs.

Build a Capacity Management Infrastructure

Figure 7 illustrates our capacity management infrastructure, which consists of three main layers:

- A foundation of business intelligence applications that provide data integration analytics.
- An automated workflow layer that addresses planning, allocation, purchasing, and provisioning.
- A set of integrated processes that can identify capacity issues, help with decision making, and request computing resources based on demand signals.

Define Capacity Planning Key Performance Indicators

We use the following key performance indicators (KPIs) to evaluate how well our capacity planning process works and where we can improve:

- Unforeseen purchases—the number of deviations from the capacity plan and total dollars spent on unplanned capacity expenditures.
- Number of capacity incidents.
- Planned and actual variance improvement, measured by a period-over-period change.
- Percentage of time we are unable to scale out in order to meet demand based on too little buffer.

Use Existing Supply Chain Expertise to Enable Cloud Capacity Methods

Many of the concepts in supply chain management correlate to our private cloud environment:

- Understanding the inventory buffer and minimizing it to a level that allows us to meet demand without having excessive waste in overhead.
- Understanding all relevant supply resources associated with successful scale out. We analyze the storage, network IP ranges, compute, memory, and the monitoring environment capacity to determine when they are approaching maximum usage and how long it will take to replenish them, which ones are part of the critical path for provisioning, which ones require dual sourcing, and so on.
- Understanding demand, such as determining whether it is cyclical in nature or a pure linear increase. Recognizing downtrends is also important to better optimize the demand-to-supply ratio.

We built on our existing knowledge of supply chain management techniques as we developed our capacity management processes and infrastructure.

SUPPORTING ON-DEMAND SELF-SERVICE

On-demand self-service is a key measure of success for our cloud initiative; however, without underlying business logic, controls, and transparency, an unconstrained on-demand enterprise private cloud will quickly exceed capacity by doling out allocations beyond supply.

We used the experience and capabilities from previous automation efforts in our Design computing environment to develop an on–demand self–service portal, which represents an enormous leap in data center agility. By instituting a hosting automation framework that includes entitlement, quotas, transparent measured services, and data-driven business logic, we are establishing a true enterprise private cloud that provides a consumer-focused self-service portal that improves business agility and velocity.[3] Currently, 80 percent of new server requests in the Office and Enterprise environment use our self-service portal.

Through on-demand self-service provisioning, we reduced the time to deploy new infrastructure services from 90 days to an SLA of less than three hours. Many of the requests are provisioned within 45 minutes. We will continue to enhance and extend this solution across the enterprise with the goal of provisioning services in just a few minutes. This is the first step in supporting our end goal of enabling Intel's engineers to obtain in less than a day the right compute infrastructure to begin working on an innovative idea.

Automate the Existing Enterprise and Transform Business Processes

To establish a VM time-to-provisioning goal of less than three hours for 80 percent of our Enterprise computing environment requires a high degree of business process automation so that customer-initiated provisioning of IT services does not increase the support burden for IT Operations or impact other customer environments on shared infrastructure.

Transforming business processes was the biggest barrier to achieving the benefits of our on-demand self-service implementation, with manual controls impeding change and agility. The number of control points, double-checks, and handoffs between teams created a very cumbersome process, which remained static as business needs evolved. Documenting business processes, discovering valuable lessons,

and repeating best practices allow us to move away from the as-is process to a new paradigm of self-service.

At the core of our self-service functionality are Web services for receiving and responding to service requests, a database to track the status and progress of these requests, a scheduler to fulfill requests, and an orchestration engine with a set of workflows to complete the tasks.

Continually Assess Opportunities to Accelerate or Extend Cloud Offerings

Workflow automation is an extremely important component of cloud computing, and is a quickly evolving field. Therefore, we need to keep up-to-date on current products and technologies.

The cloud environment is quickly maturing to match more closely the capabilities we have used in grid computing for many years. We expect that even by the time this paper is published there will have been a number of new startups as well as significant advances in existing solutions that will make cloud automation more viable and readily accessible.

Use a Phased Approach and Extend Capability Iteratively

Approaching the entire project in phases lets us build best-in-class solutions, and then extend those solutions to the next level. Table 2 shows the roadmap for our self-service provisioning.

Our strategy has been to release automation in phases to validate functionality and new architecture prior to releasing it across the business. In Phase 1, we released self–service only to the virtual development environments and incorporated measured services only for IT Operations. Operations-facing automation allows IT to provision VMs and helps verify that business processes and products are mature before extending these capabilities broadly. In Phase 2, we expanded self-service to production virtual servers and exposed measured services data to users. Currently, our portal is used for 80 percent of all VM requests.

RESULTS

As Intel expands into new consumer and business services that create revenue sources and help drive demand for Intel® processors, Intel IT supports this strategy by providing value-added solutions as well as scalable, on-demand hosting capacity.

Additionally:

- Our cloud comprises 40 percent of production workloads—not just development environments.
- We support real-world business applications across engineering, Human Resources, ERP, Finance, and more.
- We have enabled mission-critical and externally facing applications.
- More than 60 percent of our Office and Enterprise environment is virtualized today—a 3x increase during 2010.
- About 80 percent of requests for cloud services come from the self-service portal as opposed to manually.

NEXT STEPS

We are now extending the value of the cloud to more groups across Intel—and we are further accelerating delivery of new services. To quickly expand capacity that enables Intel's services businesses to meet unpredictable spikes in demand, we are adding rapid, elastic scaling for Web-based applications.

We have laid the foundation to use hybrid clouds to further increase this scalability and provide burst capacity. This year, we are sharing capacity across multiple resource pools; in 2012 we plan to share capacity across data centers and then expand to hybrid use of secure external clouds.

In addition, we intend to accomplish interim goals, including:

- **Design for failure.** As specific components in the end-to-end service fail, automated remediation will fix 95 percent of situations. In addition, nodes and components will be deployed and managed through automation in a way that allows for zero business impact from IT infrastructure downtime.
- **Broaden our cloud offerings.** We can apply what we have learned from our experience with IaaS to providing SaaS, thereby achieving even greater cost and productivity benefits.

Table 2. Development of On-Demand Self-Service Provisioning at Intel

2009	2010	2011	2012
▪ Intel Design grid automation ▪ Self-service automation for development consumers	▪ Self-service automation for internal production consumers ▪ Operator-facing automation for internal environment	▪ Operator-facing automation for demilitarized zone (DMZ) environments ▪ VM collections, rapid scale-out (elasticity & agility) ▪ Self-service automation for our DMZ for limited non IT users	▪ End-to-end cloud service monitoring ▪ Self-service automation for our DMZ for all with business need ▪ Platform as a service (PaaS) self-service

- **Enhance the self-service portal to include control panel functionality.** In the future, our customers will be able to define and manage multi-tiered application and service infrastructure with a single request. In addition, we want to extend existing portal functionality so that it supports provisioning, monitoring, configuration, patching, and backup. Also, the interface eventually will provide transparency into consumption performance to SLA attributes for the customer's provisioned and managed services.
- **Extend automation.** The automation and business logic will decide what type of services are appropriate, based on business requirements, security classification, workload characteristics, and available capacity. The selection of the service location will also be automatic—public, private, or hybrid cloud. Workloads will be dynamically migrated to higher performance infrastructure and back as demands change through an application's life cycle. All components and nodes will be dynamically added or removed immediately as necessary. Key service-level objectives will be available to the consumer, such as availability, performance, and configuration compliance.
- **Use a combination of internally hosted and Internet-hosted monitoring solutions.** Building on successful PoCs, we intend to integrate our end-to-end health monitoring capability across all of our services domains. Our goal is to be able to handle automated remediation regardless of where the service is hosted.

CONCLUSION

Although our private cloud efforts are not complete, we have delivered significant accomplishments, primarily in the area of supplying compute infrastructure through self-service provisioning:

- Virtualization of approximately 60 percent of our computing environment across Design, Office, Manufacturing, Enterprise, and Services, with a goal of 75 percent.
- A reduction of server provisioning time from 90 days to three hours. Our goal is to further reduce this to a matter of minutes.
- Self-service on-demand handling of 80 percent of server provisioning requests in the Office and Enterprise environment.

These results support our business goals of achieving an 80-percent effective utilization of IT assets, a consistent increase in business velocity, and zero business impact from IT infrastructure downtime.

As we move into 2012, our focus is shifting to supplying data and application platform services through the cloud. During the past two years, we have identified best practices in several areas that have helped us maximize the business benefit of cloud computing. We will continue to apply these best practices as we finish developing an enterprise private cloud that delivers a highly available computing environment featuring highly secure services and data on-demand to authenticated users and devices from a shared, elastic, and multitenant infrastructure. Beyond that, these best practices will also help us reach our strategic goal of a robust hybrid cloud environment.

ACRONYMS

CRM	customer relationship management
DMZ	demilitarized zone
DOMES	Design, Office, Manufacturing, Enterprise, and Services
ERP	enterprise resource planning
IaaS	infrastructure as a service
iMODS	Infrastructure Management Operational Data Store
ITIL	Information Technology Infrastructure Library*
KPI	key performance indicator
LUN	logical unit number
NPV	net present value
PaaS	platform as a service
PoC	proof of concept
PRD	project requirements document
PVLAN	private virtual LAN
SaaS	software as a service
SAN	storage area network
SIZ	secure internal zone
SLA	service-level agreement
TTM	time to market
VM	virtual machine

For more information on Intel IT best practices, visit www.intel.com/it.

[1] For more information, see "Applying Factory Principles to Accelerate Enterprise Virtualization." Intel Corporation, February 2011.

[2] For more information, see "Overcoming Security Challenges to Achieve Pervasive Virtualization," Intel Corporation, November 2011.

[3] For more information, see "Implementing On-Demand Services Inside the Intel IT Private Cloud." Intel Corporation, October 2010.

Printed in USA
1211/ABC/KC/PDF

Please Recycle
326182-001US