MOTIVATION TO LEVERAGE CLOUD AND SERVICE GRID TECHNOLOGIES

PAIN POINTS THAT CLOUDS AND SERVICE GRIDS ADDRESS

THOMAS B WINANS AND JOHN SEELY BROWN

MAY 2009

Introduction

It was September 2008 when Larry Ellison was asked about whether or not Oracle would pursue a cloud vision. His response at the time was that Oracle would eventually offer cloud computing products. But in the same conversation, Mr. Ellison also noted - in his inimitable fashion - that cloud computing was such a commonly used term as to be "*idiocy*".

Fair comment, actually, given that cloud computing *is* such an overloaded term. It can be used as a synonym for *outsourced data center hosting*. It can be used to define what Salesforce.com and Netsuite do – they offer *software as a service*. Some have referred to the Internet as *the cloud …* an uber cloud containing all others. Cloud computing sometimes is imprecisely used to reference *grid computing* for database resource management or massively parallel scientific computing. And cloud computing has been taken to mean *time sharing*, which is both a style of business and a technology strategy that sought to share expensive computing resources across corporate boundaries at attractive price points. It appears *on the surface* that the IT industry *has* redefined much of what it does now *and has done for quite a while* as cloud computing, and that Oracle indeed might need only to change verbiage in a few of its ads to align them with a well formed cloud vision.

But today's IT leaders are operating in a business climate in which intense commoditization and change force deployment of *new* IT-enabled business processes and require acknowledgement that business processes and architectures that are fixed/rigid in their definition will not scale to large networks of practice, that IT budgets may have reached the point where conventional internal cost cutting can wring out only nominal additional value unless business and IT processes change, and that *doing business internationally* is *not* the same as *conducting global business* (i.e., outsourcing is not equivalent to organizing and conducting global business) So, while Mr. Ellison's remarks may express the sentiments of many IT leaders today who have spent a considerable amount on infrastructure, they *cannot* be correct *unless* the processes and techniques developed in IT during the past 20 years *will be the foundation for processes and techniques of the next 20 years*.

We do not believe that 20th century IT thinking can or should be the defacto foundation for 21st century IT practices. We believe that now, more than ever before, IT matters, and it has already become the critical center of business operations today. As such, IT leaders have no choice but to continue to chase cost and margin optimization. They *also* have no choice but to carefully set and navigate a course to rennovate and/or replace 20th century practices with 21st century practices and technologies so that product lines and services that companies offer today can remain relevant through significant market transitions.

This paper is the first in a set of three that attempt to establish a thought framework around cloud computing, its architectural implications, and migration from current computing architectures and hosting capabilities to cloud computing architectures and hosting services. We begin by exploring three IT pain points that can be addressed by cloud and service grid computing. Subsequent papers more completely elaborate these pain points and methods to handle them.

IT Pain Points

There probably is a very large number of *IT Pain Points* that IT leaders in today's corporations would want to see addressed by cloud and service grid computing. We highlight three in this paper:

- Data Center Management;
- Architecture Transformation and Evolution (evolving current architectures or beginning from scratch); and
- Policy-based Management of IT Platforms.

Pain point: Data Center Management

Summary: The challenges of managing a data center, including network management, hardware acquisition and currency, energy efficiency, and scalability with business demands, all are costly to implement in a way that easily expands and contracts as a function of demand. Also as businesses aggregate and collaborate in global contexts, data center scalability is constrained by cost, the ability to efficiently manage environments and satisfy regulatory requirements, and geographic limitations. Adding to this complexity the need to manage change demanded by today's changing global and corporate climates underscores the need for transforming the ways that we manage data centers: methods of the past 5-10 years do not scale and do not provide the requiste agility that is now needed.

Cloud solutions: Cloud solutions can form the basis of a next generation data center strategy, bringing agility, elasticity, and economic viability to the fore:

- Affordable elasticity, scalability
    - Resource optimization through virtualization
    - Management dashboards simplify responses evoked by seasonal peak utilization demands or business expansion
    - Finer-grained container management capabilities (i.e., like Cassatt's) will serve to fine tune elasticity policies
    - Capability to affordably deploy many current technology-based applications as they exist today, possibly to re-architect them over time
    - Miminized capital outlay, which is especially important for startups where initial funding is way too limited to use to capitalize infrastructure
    - Extreme elasticity – handling spikes of traffic stemming from something catching on or 'going viral' (e.g., having to scale from 50 to 5000 servers in one day because of the power of social media)
- Affordable and alternative provisioning of disaster recovery
    - Cloud data storage schemes provide a different way to persistently store certain types of information that make explicit data replication unnecessary (storage is distributed/federated transparently)
    - Creation of virtual images of an entire technology stack streamlines recovery in the event of server failure

- Utility computing management platforms enable consistent management across data center boundaries. Evolution of utility computing to enable cloud composition will simplify implementation of failover strategies

- Affordable state-of-the-art technology
  - The exponential nature of digital hardware advances makes the challenge of keeping hardware current particularly vexing. Buying cloud services transfers the need to keep hardware current to the cloud vendor – except, possibly, as this applies to mainframe or other legacy technologies that remain viable
    - It is reasonable to expect cloud vendors to offer specialized hardware capabilities (e.g., DSP/GPU/SMP capabilities) over time in support of gaming/graphics, parallel/multi-threaded applications, etc.
    - Specialized hardware needs (e.g., mainframe-based) probably will not be the responsibility of the cloud vendor, but there is no reason why a private cloud/service grid should not be able to be composed with a public cloud/service grid

- Operational agility and efficiency – cloud vendors will oversee management of hardware and network assets at a minimum. Where they provide software assets or provision a service grid ecosystem, they likely will provide software stack management as well
  - Management of assets deployed into a cloud is standardized. Management dashboards simplify the management and deployment of both hardware and software

- Energy efficiency becomes the cloud vendor's challenge. The scale of a cloud may well precipitate the move to alternative cooling strategies (e.g., water vs. fan at hardware (board/hardware module) levels), air and water cooling of data centers, increased management software capabilities to interoperate with data center policies to control power up/down of resources and consolidate (on fewer boxes) processes running in a cloud given the visibility to utilization, SLAs, etc. One could even imagine implementing a power strategy that continuously moves resource intensive applications to run where power is less expensive (e.g., power might be less expensive at night than the day so keep run this application on the dark side of the earth)

- Big enterprise capabilities for small company prices
  - Startups can use and stay with cloud solutions as they grow and become more established

- Hardware and data center cost savings and staff cost optimizations enable businesses to self fund innovative IT initiatives
  - Those who wish to leverage the cloud's functional capabilities will have to build their own capabilities (e.g., services and/or applications) to interoperate with resources in the cloud provided that they wish to do more than simply use a cloud as outsourced hosting

- Security compliance (e.g., security of information and data center practices, PCI data security compliance) will increasingly become the responsibility of cloud vendors

- This will be true especially as clouds become/evolve into service grids, as cloud vendors geographically distribute their capabilities, as as specialized clouds are provisioned to serve specific industries

Pain point: Architecture Transformation/Evolution (The Brownfield vs. Greenfield Conundrum)

Summary: Significant investment in application platforms in the last 10 years have resulted in heterogeneous best-of-breed application systems that have proved hard and costly to integrate *within corporate boundaries*. Scaling them *beyond* corporate boundaries into corporate networks of practice takes integration to a level of complexity that appears insurmountable, but the perceived costs of starting fresh seem equally so. IT leadership knows it must do something about the application portfolio it has assembled to make it interoperable with partner networks without requiring massive technology restructure in unrealistically short time periods. It also knows the business must quickly respond to global market opportunities when they present themselves. How does IT Leadership guide the architectural evolution and transformation of what exists today to enable rapid fire response without starting from scratch or tring to change its application platform in unrealistic time periods?

Cloud solutions: Cloud solutions can form the basis of an application portfolio management strategy that can be used to address tactical short term needs, e.g., interoperability within a business community of practice using the cloud to provision community resources, and to address the longer term needs to optimize the application portfolio and possibly re-architect it.

- Cloud vendors offer the capability to construct customized virtualized images that can contain software for which a corporation has licenses. Hosting current infrastructure in a cloud (where such is possible) provides an isolated area in which a corporation or corporate partners (probably on a smaller scale due to integration complexities associated with older infrastructure and application technologies) could interoperate using existing technologies
  - Why would companies do this?
    - To move quickly with current platforms
    - To economically host applications, minimize private data center modifications, and, in so doing, self fund portfolio optimization and/or re-architecture work
    - Use current capabilities, but shadow them with new capabilities as they are developed – ultimately replacing new with old
    - Simplify architecture by removing unnecessary moving parts
- Cloud vendors offer application functionality that could replace existing capabilities (e.g., small-to-large ERP, CRM systems, etc.). Incorporating this functionality into an existing application portfolio leads to incremental re-architecture of application integrations using newer technologies and techniques (Brownfield), which, in turn, should result in service oriented interfaces that can become foundational to future state. An incremental move toward a re-architected platform hosted using cloud technologies may prove to be the only way to mitigate risks of architectural transformation while keeping corporate business running. Conversely, clouds also represent locations where Greenfield efforts can be hosted. Greenfield efforts are not

as risky as they sound given the maturity (now) of hosted platforms like SalesForce, Netsuite, etc.

  - How quickly can transformation of an existing platform be accomplished? This depends upon the architectural complexity of what is to be transformed or replaced. A very complex and tightly coupled architecture might require several years to decouple so that new architecture components could be added – assuming no Greenfield scenario is desired or feasible – whereas it might be possible to move a simply structured web application in a matter of hours. A platform that has specialized hardware requirements (e.g., there is mainframe dependency, or digital signal processing hardware is required) might have to be privately hosted, or be hosted partly in public and partly in private clouds

- Cloud APIs, together with the concepts of distribution, federation, and services that are *baked in*, provide a foundation on which to implement loosely coupled, service oriented architectures and can logically lead to better architecture
  - Web services, reliable queuing, distributed storage (non-relational with somewhat relational interfaces, and relational), provide foundational infrastructure capabilities to implement modern architectures using standardized APIs (e.g., WS-*)
  - Standardized interfaces, loose architecture couplings, and standardized deployment environments and methods increase reuse potential by making it easier to compose new services with existing services
- Clouds provide a means to deal with heterogeneity
  - Initially heterogeneity is dealt with through management layers
  - Better architecture as noted above further enhances this as heterogeneity is encapsulated beneath standardized and service oriented APIs
  - Once heterogeneity is contained, a portfolio optimization/modernization strategy can be put into place and implemented

Pain point: Policy-based Management of IT Platforms

Summary: Policy constraints are difficult to impossible to implement especially in a rapidly changing world/context. Business processes and policies are embedded in monolithic applications that comprise corporate business platforms today. Even the bespoke applications constructed in the past 10 years share this characteristic since policy and process were not treated as formal architecture components. Consequently, application scalability vis-à-vis provisioning policy driven business capabilities is limited. Organizational model changes, e.g., mergers and acquisitions (or divestitures) or corporate globalization into very loosely coupled business networks, underscore the need for policy to be made explicit. The ability to conduct business in a quickly changing world will be a direct function of the capability to manage using policy.

Service grid solutions: In one sense, this pain point can be considered to be related to the The Brownfield/Greenfield Conundrum that, if addressed, results in a distributed/federated, service oriented, loosely coupled architecture in which policy *can* be factored out and considered a first order architecture component. However, it also is clear that: (1) *everyone* does not need policy

factored like this; and (2) where policy must be exposed may vary by domain, community, geography, etc. Hence we deal with this pain point separately and consider the need to address it a prerequisite condition to leveraging the full capabilities of a service grid.

Corporations require enterprise qualities in architectures, and they will have the same expectations of clouds where they will deploy critical platforms. Scaling architectures for use in increasingly larger communities as well as simply making platforms that are far more configurable and compositional requires the ability to expose policy extension points that can be incorporated into a management scheme that can implement and enforce policy across the entire technology stack. The result of such architecturally pervasive policy management is that control over environmental as well as business constraints is provisioned. When corporate architectures are deployed into service grids, these policy extension points must be used even in grids composed of other grids.

- Cloud vendors are implementing *utility computing* management capabilities which are, themselves, policy driven. As these capabilities are further refined, it will become feasible to integrate business policies with infrastructure management policies on which business and infrastructure service level agreements/processes can be based
- Outside-In architectures (in our view, service oriented architecture properly done) provision interfaces that easily align with business processes and minimize architecture complexity, resulting in a simpler architecture in which policy is externalized. Externalized policy provides the opportunity for business policy to join with infrastructure policy
- Externalized policy provides the foundation on which policy-driven business processes can be constructed and managed. This results in increased business agility because it simplifies how businesses interoperate: they interoperate at the business process level, and not at the technology level; policies can be changed with significantly less impact on the code that provisions business functionality
- Cloud vendors use containers to deploy functionality. As these containers become permeable such that their contents can both be managed and expose policy extension points, then policies can span the entire cloud and grid technology stacks
- Service grids provision architecture components, e.g., policy engines and interaction services, that enable policy to be managed/harmonized explicitly and separately from other business functionality – across architecture layers, across business networks of practice - and used as the foundation of business interactions
    - It is important to note that policy is viewed as a constraint continuum covering infrastructure management to domain (regulatory, industry/market sector) policy constraints

Concluding Remarks: To the 21st Century and Beyond

We see, in this paper, that cloud computing can be used to address tactical problems with which IT continually deals, like resource availability and reliability, data center costs, and operational process standardization. These near term objectives represent sufficient justification for companies to use cloud computing technologies even when they have no need to improve their platforms or practices. But there are longer term business imperatives as well, like the need for a company to be agile in combining their capabilities with those of their partners by creating a

distributed platform that will drive aggressively toward cloud and service grid computing. We believe that clouds, service grids, and service oriented architectures are technologies that will be fundamental to 21st century corporations' successfully navigating the changes that they now face.

The pain points discussed above illustrate a progression of change that most corporations have already begun, whether they are just starting up or are well established. We began with use of cloud hosting services as an alternative to self hosting, or as an alternative to other current day 3rd party hosting arrangements that do not offer at least the potential of cloud computing. For those companies that need to pursue implementation and management of a service oriented architecture, we discussed pain points relating to re-architecting current platforms to leverage cloud computing, and the possible need to formalize the way that policy is used to manage IT platforms within and across service grid boundaries.

Many of the concepts mentioned in the pain point discussions are architectural, and are not defined at all in this paper. However, they are more completely elaborated in our other papers, called *Demystifying Clouds: Exploring Cloud and Service Grid Architectures*, and *Moving Information Technology Platforms To The Clouds: Insights Into IT Platform Architecture Transformation*.

About The Authors

Thomas B (Tom) Winans is the principal consultant of Concentrum Inc., a professional software engineering and technology diligence consultancy. His client base includes Warburg Pincus, LLC and the Deloitte Center for the Edge. Tom may be reached through his website at http://www.concentrum.com.

John Seely Brown is the independent co-chairman of the Deloitte Center for the Edge where he and his Deloitte colleagues explore what executives can learn from innovation emerging on various edges, including the edges of institutions, markets, geographies and generations. He is also a Visiting Scholar and Advisor to the Provost at USC. His web site is at http://www.johnseelybrown.com.