There has been such great hope around the pay-for-what-you-use savings of cloud economics. But real cloud economics are only achieved by dynamically optimizing the balance between operational and financial objectives of using the cloud. Simply over-provisioning until an intuitive feeling of comfort and safety is achieved is no more than fantasy cloud economics, and to achieve “perfect provisioning” companies need highly intelligent tools which can not only iteratively learn resource usage patterns, but can provide actionable recommendations for how to optimize cloud deployment by rightsizing and choosing the most economical pricing plans.
Understandably, the enticing combination of agility, scalability and cost savings has attached a great deal of hope to the economically favorable potential of cloud computing. In 2011, however it became evident that the math of cloud economics might still be somewhat in need of some refinement. It turns out that cloud resource deployments can’t always be characterized or allocated using simple models – and the complexity of cloud pricing and provisioning is quickly moving beyond human capacity to comprehend, much less act upon.

For many applications, ensuring operational performance (SLA-type performance assurances) means choosing to over-provision cloud resources with enough of a safety margin to ensure performance requirements are always met. This is colloquially known as the “so-I-can-sleep-at-night” approach to cloud provisioning.

It has been historically impossible to know exactly how much safety margin is required at any one point in time, much less in a variable fashion as resource demands change dynamically. A common choice, therefore, has been to over-provision cloud computing resources by excessive amounts in order to be absolutely certain that performance demands will be met.

Over-provisioning of cloud resources has, in the absence of other choices, become an epidemic. The result is that many organizations are investing in cloud resources they simply do not use. Unused resources produce, of course, a return on investment (ROI) of exactly zero.

Cost versus Performance

In introducing the April 2011 research report, “The Three Stages Of Cloud Economics,” James Staten, Vice President and Principal Analyst at Forrester Research, Inc. wrote, “We all believe that cloud computing can save us money but aren’t sure how to maximize these savings.”

Staten’s question is the essence of cloud economics – finding the perfect level of provisioning, at any particular moment in time, that optimizes the balance between the operational and financial objectives of using the cloud.

This constant tension of cost versus performance has always existed in IT resource provisioning. But the lack of clarity in cloud deployments has made it more difficult than ever to achieve the optimal balance.
What are cloud economics?

**Hopeful (fantasy) cloud economics** = Because I only pay for what I use, going to the cloud automatically saves my company a lot of money.

**The barrier to fantasy cloud economics** = The risk of compromised performance requires me to *over-provision*.

**Real cloud economics** = the cost of actual resources required by my cloud deployment, including a *safety-margin* that *ensures* my cloud resources meet my company’s performance demands.

**Real cloud economics are only achieved by dynamically optimizing the balance between operational and financial objectives of using the cloud.**

"The key to cloud economics is to use cloud computing optimally."

*"The Three Stages Of Cloud Economics"*  
Forrester Research, April, 2011 – by James Staten

The Forrester report also stated: “The key to cloud economics is to use cloud computing optimally.” Certainly every company who has ever moved to the cloud understands this – but the question of how to reach optimal deployment, especially dynamically, remains.

Many applications running on the cloud need to be available 24/7, while guaranteeing a smooth and consistent customer experience. This must be the case, no matter how many users are accessing the application or how dynamic application usage becomes. For example during the holiday shopping season, e-commerce traffic rockets just when e-commerce companies cannot afford any downtime or sluggish response times. Dynamic demands can introduce erratic variations in workloads, servers, storage, databases, and network access needs.

A company’s operational performance must be protected in all cases, so users of cloud resources inside or outside the company never feel a reduction in performance quality – they never feel that “going to the cloud” has introduced a performance compromise.
When provisioning cloud resources, there are two kinds of risk at stake:

**Operational risk** – Insufficient provisioning of cloud computing resources risks service-level performance degradation.

**Financial risk** – Every amount of safe-margin over-provisioning of cloud resources represents unused resources which represents a zero-ROI expense.

Stating the obvious, no organization can accept the operational risk of compromising service level objectives. Some level of over-provisioning of cloud resources, with enough safe-margin headroom to guarantee an organization’s needs are met is therefore necessary – the only path is to take some level of financial risk. But how much investment in a performance safety margin is enough to mitigate risk, and at what point does that safe-margin investment cross over into being over-provisioning – an unnecessary over-expenditure?

Simply over-provisioning until an intuitive feeling of comfort and safety is achieved is no more than *fantasy* cloud economics. *Real* cloud economics are achieve with optimal, “perfect provisioning” of cloud resources. And to achieve perfect provisioning, companies need highly intelligent tools which can not only iteratively learn resource usage patterns, but can provide actionable recommendations for how to optimize cloud deployment by rightsizing and choosing the most economical pricing plans.

In many ways, “Cloud” is functionally synonymous with today’s highly desirable business attribute of “agility.” One result is that new cloud vendors and services appear every day, each offering different cloud configurations and pricing plans that change with breathtaking frequency.
Cloud vendors do typically provide the flexibility to select an instance type from among a menu of many to help companies aim for perfect provisioning—instance types are variable in terms of compute units (CPU cycles), memory, and I/O capacity. This would be enough in a perfect world, but in practice, cloud computing resources run in shared environments, so actual performance of a particular instance is affected by other workloads running simultaneously on the same hardware. Selecting larger instance types certainly may provide the desired performance requirements for longer periods of time, but at the cost of greater investment.

There are many different pricing plans, such as reservations, on-demand and the spot market, each of which is appropriate for different resources utilization requirements.

Using the reservation model, for example, cloud vendors permit the reserving of resources for various engagement periods to gain significant cost savings. Reservation is a purely administrative operation, and no instance modification is required.

Both selecting the appropriate instance type and determining the optimal reservation size actually depends on understanding desired application-performance targets.

“Merely measuring something has an uncanny tendency to improve it. Pretty soon you start noticing what makes the number go up, and you start to do more of that.”

W. Edwards Demming, Father of Total Quality Management (TQM), whose fundamental principal was to strive for continuous improvement.

Pricing plan revisions and reductions are introduced constantly, making it challenging to keep track of the latest best-price updates. AWS, for example, changed and reduced their pricing schedule 8 times in 2011. Just a few weeks into 2012, both AWS and Rackspace each cut their fees, yet again.

These choices also require going through an intelligent, recursive process of planning, tracking (deep analysis of usage patterns) and optimization to get to real cloud economics—the desired performance at the lowest cost.

Enabling Real Cloud Economics

Cloud economics must be approached with a methodical and iterative cycle of deployment planning, performance/expense tracking, and resource/expense optimization. This is easier
said than done, of course, and optimization must be able to circle back to deployment planning, even as optimization is driven by tracking of past performance requirements plus a predictive understanding of future performance demands.

Clarity, Visibility, Measurement

It is fortunate for IT managers that a number of companies have recently come to the market with tools for monitoring usage of cloud resources.

Cloud resource monitoring is crucial for measuring a company’s consumption and spending on cloud computing resources. As the adoption of cloud services accelerates, companies need easy, accurate and dynamic tools to help them keep track of how much they’re spending and on what. This is the only way to effect meaningful change and therefore savings in the cloud. To achieve, in other words, real cloud economics.

The value of monitoring is infinite (since you can’t optimize without it), but is also quickly become a check-the-box requirement for anyone deploying cloud resources – simply monitoring is headed toward commodity status.

Understanding cloud resources in an organized, easily visualized and comprehensible fashion is the baseline of computer-based analysis.

Analytics are Not Enough

Revealing what has been provisioned in a cloud computing deployment is an incomplete first step in the process of arriving at what is needed to optimize the use and cost of cloud resources. Optimizing through trial and error, even measured trial and error, isn’t good enough anymore and it isn’t responsible behavior when intelligent tools are available to help do a better job of balancing performance and cost.

The real intelligence is in computer-based decision support that not only compiles all the information about your cloud resources, but understands better than a human can, how to actively respond to the detailed characterization of a cloud deployment. This is even more true as cloud deployments scale out to massive sizes.

Intelligent heuristics add real value to the process of achieving real cloud economics.
Decision support systems that provide actionable recommendations for optimizing cloud provisioning are what bridge the gap between real cloud economics and fantasy cloud economics – to get much closer to the “going to the cloud saves a company lots of money” dream.

Value Added Intelligent Heuristics

Cloud resources are hugely complex systems which require highly intelligent decision making in order to adjust resource allocations to save money without adversely affecting operational performance.

“Cloudyn lets users optimize spending while maintaining performance, even as the environment changes.”

Paul Burns, Neovise Research

Cloudyn has developed just such intelligent decision support heuristics – literally experience-based techniques for problem solving, learning, and discovery, applied to cloud provisioning optimization. Cloudyn’s heuristics evolve and adapt to changing conditions and requirements, and to the availability of cloud emerging technologies, services, and pricing models.

“Many organizations that deploy applications on public clouds have been surprised with unexpectedly high bills,” according to cloud industry analyst Paul Burns of Neovise Research. “Unfortunately, cost calculators, monitoring tools and rigid spending controls fall short when it comes to balancing spending and performance, Cloudyn moves beyond those approaches by providing intelligent and actionable recommendations for rightsizing cloud resources. By providing feedback on an ongoing basis, Cloudyn lets users optimize spending while maintaining performance, even as the environment changes.”

Cloudyn provides companies not just with clarity, but with real certainty, determinism, and action around cloud economics. The result is intelligent help in getting past trial & error provisioning and/or massive (and expensive) over-provisioning to avoid performance failures.

Cloudyn’s ability to monitor information and then process it, replaces the old manual trial and error process with an automated actionable approach based on prescriptive guidance that optimizes cloud usage and achieves cost reductions for a targeted performance level.

It’s true, cloud computing can save you money. But it is only true with intelligent and actionable recommendations for rightsizing cloud resources. This, and only this, will realize the hope of real cloud economics.
Your checklist for maximizing cloud investments:

• Clearly see and understand the way the company uses its cloud resources over time
• Dynamically quantify the optimal provisioning safety-margin, and rightsize accordingly
• Stay up-to-date with changing pricing plans and rates
• Continuously choose the relevant price plan to match application needs
• Include cloud economics metrics when engaging in capacity planning for new cloud applications
• Use the most intelligent and actionable tools available to aid your cloud resource management process

For additional best practices in leveraging real cloud economics, and to participate in discussions on the topic, visit www.cloudyn.com/cloudeconomics and join our Cloud Economics group on LinkedIn.