# Cloud Provider High Availability

## A Survey of Guarantees, Industry Views, & Performance

Lisa Spainhower
Tavira, Portugal
January 18, 2013

# Cloud HA Guarantee Delivery

- Availability % guarantees in public domain SLAs
- Uncommon in enterprise computing
  - Legal concerns
  - SLAs typically customized and confidential
- Characteristics of Cloud HA guarantees
  - Similar but not standard terminology
  - Various degrees of legalese
  - Little/no room for customization

# Cloud HA Guarantee Evaluation

- SLA comparison
  - Focus on market leaders

SLA components

  - HA % guarantee
    - Time period
    - Threshold
    - Scope
  - Restitution
  - Exclusions
  - Responsibilities

# Cloud HA Guarantee Comparison

| | Amazon EC2 | Azure Compute | Google Apps | Rackspace | Terremark/ Verizon |
|---|---|---|---|---|---|
| **HA Guarantee** | 99.95% | 99.95% | 99.9% | 100%(power, HVAC network) | 100% |
| **Time Period** | 365 days | Month | Month | Month | Month |
| **Threshold** | 5 minutes | 2 minutes | None | None, see below | 15 minutes |
| **Scope** | >1 Availability Zone | >1 Update Domain | Google applications | Virtual Machine | Data Center |
| **Other comments** | If usage <  365 days,  days prior to your use  will be deemed to have had 100% Region Availability | | | Server host fails, restoration or repair within 1 hour. Required server migration complete within three hours | |

# SLA Exclusions

| | Amazon EC2 | Azure Compute | Google Apps | Rackspace | Terremark/ Verizon |
|---|---|---|---|---|---|
| **Force Majeure** | Yes | Yes | Yes | Yes | Yes |
| **Third Party** | Yes | Yes | Yes | Yes | Yes |
| **Hacking, DOS, virus** | Yes | Yes | Yes | Yes | Yes |
| **Customer Invoked** | Yes | Yes | Yes | Yes | Yes |
| **Internet** | Yes | Yes | Yes | Yes | Yes |
| **Scheduled Maintenance** | Unclear | Yes | No | Yes* | Unclear |
| **Other comments** | These exclusions are primarily boilerplate using common text for all five providers. | | | Server repair <1 hour, Migration <3 hours<br>*Not to exceed 60 min/month with at least 10 days notice | "Failures of individual functions, features, infrastructure, and network connectivity [excluded]" |

# SLA Restitution

| | Amazon EC2 | Azure Compute | Google Apps | Rackspace | Terremark/ Verizon |
|---|---|---|---|---|---|
| **Credit** | 10% if <99.95 | 10% if <99.95<br>25% if <99 | 3 days if<99.9<br>7 days if <99<br>15 days if <95 | 5-100% | $1/15 min up to 50% of bill |
| **Bill affected** | Future | Current | Current | Current | Future |
| **Credit filing window** | 30 days | 1 month | 30 days | 30 days | 30 days |
| **Other comments** | | Must report within 5 days | $ instead of service permitted | | |

- Maximum restitution is one month service
  - Most is considerably less
- Credit terms varied and confusing

# SLA Responsibilities

- For all providers the customer is responsible to:
  - Detect the failure
  - Report the failure
  - Gather/transmit documentation for any claim
  - Usually, determine when service is restored
- No apparent contractual relief for degradation
  - No performance guarantee from any provider

# Analyst Views on Current Cloud SLAs

- Their purpose is to provide a basis for post-incident legal combat
- “Take it or leave it” approach, often inadequate compensation

A very large retail customer's website crashed on Black Friday for 6 hours
**Loss = $50M USD   Compensation = $300 USD**

- Gartner called Amazon and HP cloud SLAs “practically useless”
- Often unrealistic; all infrastructures will have outages

**A complaint was filed with Advertising Standards Authority against Rackspace for 100% uptime claim, stating it was not possible.**
**In its adjudication, the ASA said: “We investigated the ad under the [Committees of Advertising Practice] Code ... but did not find a breach.”**
**“The ASA considered that consumers would understand that the claim ‘100% Uptime Guarantee’ meant they would be compensated if the Rackspace Cloud network was unavailable...and concluded the claim ... was not misleading.”**

# Recommendations for Future SLAs

- SLAs should be based on specific business needs and revolve around key business metrics
- SLAs need a meaningful and realistic penalty model not restricted to service credit

# Open Cloud Manifesto

“Dedicated to the belief that the cloud should be open”

Over 400 supporters - essentially every cloud-related company big or small

“Cloud Computing Use Cases Whitepaper” proposes wide range of standards including SLA

However, doc hasn't been updated since July 2010

# Cloud Downtime Attracts Attention

- Cloud outages tend to have very high visibility and receive media attention
- Extensive coverage of AWS US-East Region 1 outages
  - March 15, June 15, June 30, Oct 22 & Dec 24
- Google services down on October 31
  - 10% of accesses unsuccessful
  - Service restored in 6 minutes

“This massive outage – however brief – shows how tenuous our “digital lives” can be.”

- Major cloud providers highly motivated to stay up

# Top 10 Outages of 2012

- According to Data Center Knowledge
- Compiled prior to Dec. 24 Amazon EC2 Outage that brought down Netflix

1. Super Storm Sandy, Oct. 29-30
2. Go Daddy DNS, Sept. 10
3. ★ Amazon EC2 Outage, June 29-30
4. Calgary Data Center Fire, July 11
5. Australian Airport Chaos, July 1
6. Windows Azure Outage, Feb. 29
7. ★ Salesforce.com Outage, July 10
8. ★ Syrian Internet Blackout, Nov.29
9. Windows Azure Outage, July 28
10. ★ Hosting.com Outage, July 28

★

★ Cloud

# Cloud HA Evaluation 2007-2012

International Working Group on Cloud Computing Resilience

- Formed in March 2012 by Telecom ParisTech and Paris 13 University
- Conducted a study of 13 major cloud providers and the outages they have experienced, as reported, over the last five years
- Published as “Downtime Statistics of Current Cloud Solutions”
- An average of 7.5 hours unavailable per year, or ~99.9% availability
- Acknowledged that the immaturity of the report and its reliance on press reports for outage details means it should be “taken with a pinch of salt”

# IWGCCR Results 2007-2012

| | Total Outage (Hr) | Average/Yr (Hr) | Availability | Cost/Hr (USD) | Cost (USD) |
|---|---:|---:|---:|---:|---:|
| 1. Amadeus | 1 | 0.167 | 99.998% | 89,000 | 89,000 |
| 2. Facebook | 3 | 0.500 | 99.994% | 200,000 | 600,000 |
| 3. ServerBeach | 4 | 0.667 | 99.992% | 100,000 | 400,000 |
| 4. PayPal | 5 | 0.833 | 99.990% | 225,000 | 1,125,000 |
| 5. Google | 5 | 0.833 | 99.990% | 200.000 | 1,000,000 |
| 6. Yahoo | 6 | 1.000 | 99.989% | 200,000 | 1,200,000 |
| 7. Twitter | 7 | 1.167 | 99.987% | 200,000 | 1,400,000 |
| 8. Amazon | 24 | 4.000 | 99.954% | 180,000 | 4,320,000 |
| 9. Microsoft | 31 | 5.167 | 99.941% | 200,000 | 6,200,000 |
| 10. Hostway | 72 | 12.000 | 99.863% | 100,000 | 7,200,000 |
| 11. BlackBerry | 72 | 12.000 | 99.863% | 200,000 | 14,400,000 |
| 12. NaviSite | 168 | 28.000 | 99.680% | 100,000 | 16,800,000 |
| 13. OVH | 170 | 28.333 | 99.677% | 100,000 | 17,000,000 |
| **Total** | **568** | **94.667** | **99.917%** | | **71,734.000** |

# CloudHarmony 2010 Case Study

- Partnered or contracted with 38 cloud vendors
- Deployed Panopta for monitoring, outage confirmation, & availability metric calculation
  - Each outage verified by 4 geographically dispersed nodes
  - All outages >5 minutes documented and confirmed via vendor contacts and/or status pages
  - Outages due to scheduled maintenance, DoS, and self-inflicted are removed

# CloudHarmony 2010 Case Study Results

| Provider | Data Center | #/min outage | SLA | Actual |
|---|---|---|---|---|
| **AWS EC2** | US East | 0/0 | 99,5% | **100%*** |
| **AWS EC2** | US West | 0/0 | 99,5% | **100%** |
| **GoGrid** | US West | 0/0 | 100% | **100%** |
| **Linode VPS** | London | 0/0 | 99,9% | **100%** |
| **OpSource Cloud** | VA, US | 0/0 | 100% | **100%** |
| **Storm on Demand** | MI, US | 0/0 | 100% | **100%** |
| **VoxCLOUD** | EU | 0/0 | 100% | **100%** |
| **GoGrid** | US East | 1/2.3 | 100% | **99.999%** |
| **Joyent Smart Machines** | Andover, MA | 1/3 | 100% | **99.999%** |
| **VoxCLOUD** | Singapore | 1/5.5 | 100% | **99.999%** |
| **Speedyrails VPS** | Peer1 Quebec | 1/2.2 | 99,9% | **99.999%** |
| **Rackspace Cloud** | Dallas, TX | 1/8.7 | 100% | **99.998%** |
| **SoftLayer CloudLayer** | Dallas, TX | 4/13.9 | 100% | **99.997%** |
| **Hosting.com** | Colorado | 1/1.4 | 100% | **99.997%** |
| **AWS EC2** | APAC | 5/14.8 | 99,5% | **99.996%** |
| **Linode** | Atlanta | 10/26.9 | 99.9% | **99.995%** |
| **Joyent Smart Machines** | Emeryville, CA | 4/15,2 | 100% | **99.994%** |
| **Terremark vCloud** | FL, US | 7/37.9 | 100% | **99.993%** |
| **AWS EC2** | EU West | 3/36 | 99.5% | **99.993%** |
| **Speedyrails VPS** | Canix Quebec | 9/38.7 | 99.9% | **99.992%** |
| **Linode** | Fremont, CA | 13/71.9 | 99.9% | **99.986%** |
| **Zerigo** | CO, CA | 9/66.8 | 99.99% | **99.985%** |
| **SoftLayer CloudLayer** | DC, US | 31/86.7 | 100% | **99.984%** |
| **SoftLayer CloudLayer** | WA, US | 13/106.8 | 100% | **99.980%** |
| **Linode** | NJ, CA | 14/145.7 | 99,9% | **99.972%** |
| **VoxCLOUD** | NY, US | 12/146.3 | 100% | **99.972%** |
| **CloudSigma** | Switzerland | 22/59.9 | 100% | **99.972%** |
| **Hosting.com** | KY, US | 4/38.7 | 100% | **99.955%** |
| **ThePlanet Cloud Servers** | TX, US | 34/144.3 | 100% | **99.955%** |
| **Gandi VPS** | France | 4/147.7 | 99.95% | **99.955%** |
| **Linode** | Dallas | 21/258.2 | 99.9% | **99.951%** |
| **NewServers** | FL, US | 39/288.7 | 99.99% | **99.945%** |
| **VPS.NET** | UK | 8/250.3 | 100% | **99.921%** |
| **VPS.NET** | US Central | 12/342.9 | 100% | **99.892%** |
| **Flexiant** | UK | 83/820.3 | 100% | **99.844%** |
| **VPS.NET** | US West | 32/576.5 | 100% | **99.819%** |
| **ReliaCloud** | MN, US | 23/1941.5 | 100% | **99.626%** |
| **VPS.NET** | US East | 6/1224.1 | 100% | **99.616%** |

***16 of 38 meet/exceed SLA**

# Cloud Availability and Outage Reporting

- Using the same methodology, CloudHarmony continuously monitors and reports last 90 days availability % for >100 providers
- Continuous availability tracking for variable time periods (6 hr- 30 days) from CloudSleuth which also reports on recent outages

# My Observations

- Evolving business needs will drive SLA terms
  - Current users may lack even a profit plan
- Providers weak in monitoring and management
  - Unable to achieve the cloud promise of instantaneous, automatic right-sizing
- Availability is visible and quite high
  - Third party real time monitoring a plus

# BACKUP

# Open Cloud Manifesto

## Dedicated to the belief that the cloud should be open

1. Cloud providers must work together to ensure that the challenges to cloud adoption (security, integration, portability, interoperability, governance/management, metering/monitoring) are addressed through open collaboration and the appropriate use of standards.

2. Cloud providers must not use their market position to lock customers into their particular platforms and limit their choice of providers.

3. Cloud providers must use and adopt existing standards wherever appropriate. The IT industry has invested heavily in existing standards and standards organizations; there is no need to duplicate or reinvent them.

4. When new standards (or adjustments to existing standards) are needed, we must be judicious and pragmatic to avoid creating too many standards. We must ensure that standards promote innovation and do not inhibit it.

5. Any community effort around the open cloud should be driven by customer needs, not merely the technical needs of cloud providers, and should be tested or verified against real customer requirements.

6. Cloud computing standards organizations, advocacy groups, and communities should work together and stay coordinated, making sure that efforts do not conflict or overlap

# Cloud Computing Use Cases Whitepaper Minimal Recommended SLA Metrics

☐ **Throughput** – How quickly the service responds

☐ **Reliability** – How often the service is available

☐ **Load balancing** – When elasticity kicks in (e.g., new VMs are booted or terminated)

☐ **Durability** – How likely the data is to be lost

☐ **Elasticity** – The ability for a given resource to grow infinitely, with limits (the maximum amount of storage or bandwidth, for example) clearly stated

☐ **Linearity** – How a system performs as the load increases

☐ **Agility** – How quickly the provider responds as the consumer's resource load scales up and down

☐ **Automation** – What percentage of requests to the provider are handled without any human interaction

☐ **Customer service response times** – How quickly the provider responds to a service request. This refers to the human interactions required when something goes wrong with the on-demand, self-service aspects of the cloud