Load Balancing / Auto Scaling

CSCI 4417/5417

Introduction to System Administration

East Tennessee State University
Department of Computing
Jack Ramsey, Lecturer

Buzz Terms

High Availability - Critical systems must be operative as much as possible

Total Cost of Ownership (TCO) - Total cost for acquiring and maintaining a system

Buzz Terms

Acquisition Cost - Cost of acquiring the system

Research

Design

Source the produces

Purchase the products

Install the system

Develop applications

Train users

Deploy the system

Buzz Terms

Maintenance Cost - Cost of maintaining availability to end users

- Systems management
- Maintenance of hardware and software components
- User support
- Environmental factors
- Other (depending on type of system and circumstances)

Buzz Terms

While the above may seem straightforward, quantifying each cost is difficult

- Accounting
- No record of all employee activities by task
- No accurate inventory

TCO is an attempt to understand what costs could reasonably be, not exact figures

Availability is the most significant contributor to TCO

Buzz Terms

Service Level Agreement

- Availability and uptime
- Response time
- Number of concurrent users supported
- Performance benchmarks
- Advance notification of change
- Help desk response time
- Provide usage statistics

May also include a plan for addressing downtime and documentation for how the service provider will compensate customers in the event of contract breach.

Availability %	Downtime per year	Downtime per month	Downtime per week	Downtime per day
90% ("one nine")	36.5 days	72 hours	16.8 hours	2.4 hours
95%	18.25 days	36 hours	8.4 hours	1.2 hours
97%	10.96 days	21.6 hours	5.04 hours	43.2 minutes
98%	7.30 days	14.4 hours	3.36 hours	28.8 minutes
99% ("two nines")	3.65 days	7.20 hours	1.68 hours	14.4 minutes
99.8%	17.52 hours	86.23 minutes	20.16 minutes	2.88 minutes
99.9% ("three nines")	8.76 hours	43.8 minutes	10.1 minutes	1.44 minutes
99.99% ("four nines")	52.56 minutes	4.38 minutes	1.01 minutes	8.66 seconds
99.999% ("five nines")	5.26 minutes	25.9 seconds	6.05 seconds	864.3 milliseconds
99.9999% ("six nines")	31.5 seconds	2.59 seconds	604.8 milliseconds	86.4 milliseconds
99.99999% ("seven nines")	3.15 seconds	262.97 milliseconds	60.48 milliseconds	8.64 milliseconds
99.999999% ("eight nines")	315.569 milliseconds	26.297 milliseconds	6.048 milliseconds	0.864 milliseconds
99.9999999% ("nine nines")	31.5569 milliseconds	2.6297 milliseconds	0.6048 milliseconds	0.0864 milliseconds

Buzz Terms

Scalability

The ability of a system to handle spikes in loads by launching additional resources

Can be proactive or reactive

Buzz Terms

Load Balancing

Spreading the load across several - or - more machines

Different scheduling schemes

- Round robin
- Weighted round robin
- By connections
- Many others

Buzz Terms

Fault Tolerance

The way in which an operating system responds to a hardware or software failure

System's ability to allow for failures or malfunctions

- Power failure
- Using backup system
- Mirrored disks
- Multiple processors comparing data for errors
- Redundancy

References

Piedad, F. & Hawkins, M. (2001). High availability: Design, techniques, and processes. Upper Saddle River, NJ: Prentice Hall.