CAPACITY PLANNING IN THE CLOUD

Does Capacity Planning Fundamentally Change when Resources are Unlimited in the Cloud?

Alex Bewley – CTO, uptime software inc.
Why am I here?

- Been in the systems management business eighteen years
- Lots of experience with cloud
- Open Systems – so I’m from a different planet anyway
- Joined CMG in 2004
Agenda

- Growth in Cloud?
- What kinds of workloads are in the cloud?
- Explaining Amazon Web Services (AWS)
- Scaling in AWS (examples)
- Saving money is the goal
- Optimizing a growing application
Is there growth in Cloud?

![Graph showing global spending forecast by enterprises on cloud architecture from 2011 to 2017. The spending increases from $50 billion in 2011 to $250 billion in 2017.]
Who uses Cloud?

• Four kinds of companies
  – Pure Enterprise (on-premise only)
  – Pure Cloud (AWS/Azure/GCE)
  – Hybrid – cloud savvy (can manage themselves)
  – Hybrid – not savvy (can’t manage, outsource)

• What kind of Workloads
  – Always on (database, web front-end)
  – Batch - Image processing, Web crawlers, Big Data (Hadoop)
  – VDI (Virtual Desktop)
## What is AWS?

### Amazon Web Services

#### Compute & Networking
- Direct Connect
- EC2
- Route 53
- VPC
- WorkSpaces

#### Storage & Content Delivery
- CloudFront
- Glacier
- S3
- Storage Gateway

#### Database
- DynamoDB
- ElastiCache
- RDS
- Redshift

#### Analytics
- Data Pipeline
- Elastic MapReduce
- Kinesis

#### Deployment & Management
- CloudFormation
- CloudTrail
- CloudWatch
- Elastic Beanstalk
- IAM
- OpsWorks

#### App Services
- AppStream
- CloudSearch
- Elastic Transcoder
- SES
- SNS
- SQS
- SWF
How hard is it to scale in AWS?

Step 3: Configure Instance Details

Configure the instance to suit your requirements. You can launch multiple instances from the same
more.

Number of instances: 4000

Purchasing option

- Request Spot Instances

Current price

- us-east-1a: 0.0161
- us-east-1b: 0.0161
- us-east-1c: 0.240
- us-east-1d: 0.0161
- us-east-1e: 0.0161

Maximum price: $0.017
Understanding cost at this scale

- 4,000 m1.large instances (4 CPU, 7.5GB RAM, 420GB SSD)
  - @ $0.061/hour
  - $64/hr

- I provisioned these 4,000 instances in 10 mouse clicks and they came live in 7 minutes (I can do this in Virginia, Sao Paulo, Sydney, Tokyo, Singapore, Ireland, and Oregon)

- Okay, that’s nice, but what can you do with this?
Managing Crazy Capacity

• In the cloud, turning things off as fast as possible becomes the goal
  – I forget to kill my 4,000 instances
  – Monthly CPU cost is $47,656
  – Monthly storage is 420GB*4,000 = 1,680,000GB = 1.6PB
    • Storage cost is $0.05/GB-month
    • Monthly storage cost is $84,000
  – Total monthly cost for just CPU and storage is $131,656
    • A lot more than the original $64/hr that was enticing
How to save money easily

• **Telco**
  – Had 6,000 developers in AWS
  – Left instances on all the time
  – Consultant wrote a script to turn off the instances between 6pm – 8am weekdays and all weekend
    • Of 168 (24*7) hours a week, was able to shutdown instances for 13 hours a day + weekend
    • Instances only ran for 10 hrs/day weekdays – 50 hours running vs. 168 hours – only 30% the cost
Other cost leaks

• **Multi-department**
  – Lots of groups leave small instances running
  – Lingering snapshots of data
    • All the left-overs add up over time

• **No cross-department management of resources**
  – Costs are simply passed on to business groups without much global oversight
The evolution of an application in the cloud

- Turn on and leave on
- Use Reserved Instances (prepaying for discounted usage)
- Using Spot instances (heavily discounted excess capacity)
- Starting and Stopping (save state)
- Discovering queues and instance types
Big Data in the Cloud

• Example of dynamic provisioning
• Reporting servers run during the business day
• Data is dumped from SQL Server databases into flat files at night
  – SQL Servers stopped
• A Hadoop cluster of 30 spot instances is spawned to crunch the numbers (analytics)
  – These instances only exist for 4-5 hours to do the analytics
  – SQL Servers restarted
• Data reloaded into SQL Servers
  – SQL Servers stopped until business next day

• So why is this good?
  – Production reporting engines only working for business day
  – Don’t need 30 systems allocated full time – only 4-5 hours a night at really cheap rates
Summary

- Looked at how cloud “works” - AWS
- Types of applications
- Cost modeling
  - How to understand applications and make reasonable and helpful recommendations
  - How can you enable the consumption of the most amount of resource to get the job done, but in the cheapest manner possible
    - With a completely different toolset
Thank you.

alex.bewley@uptimesoftware.com
@AlexBewley