Consolidation

Ken Birman

Massive data centers

- We've discussed the emergence of massive data centers associated with web applications and cloud computing
  - Generally use web services standards
  - Run software like MapReduce (Hadoop), GFS
  - Might include tens of thousands or millions of machines
- Enormous economies of scale
  - Storage becomes remarkably cheap: 10x less than anything you can buy for "home or office use"
  - CPU cycles start to seem free

2

Can we do more?

- Companies like Google, Amazon want to
  - Keep those machines busy: If you plan to run them, the more work they can do the better
  - Migrate things onto them that generate revenue
- Amazon: leader in renting virtual machine images
  - Instead of running applications at home…
  - … why not rent a cheap machine from Amazon and run your applications there?

3

Amazon EC2

- Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud. It is designed to make web-scale computing easier for developers.
  - Web services interface
  - They give you as many virtual machines as you like… you can configure them as you wish
  - Pick the O/S and applications you need
- You pay only for resources you actually use

4

A huge success!

- Today Amazon hosts literally millions of virtual machines on EC2
  - The number actually running at any point in time varies
  - They migrate these machines within their data center to balance load and use their resources efficiently
- Thus one data center machine might host one VM image… or twenty of them
  - Idea is that you won't notice…

5

Amazon EC2: Capacity on demand

- Basic concept:
  - Virtual machines running on Amazon hardware
  - Cost is $0.10 - $0.80 per "instance hour"
    - Plus bandwidth & AMI storage
  - "Amazon Machine Image" – Xen VM with special packaging

Amazon EC2: Capacity on demand

- Basic concept:
  - AMI is stored on Amazon S3
    - Internal traffic (S3 <-> EC2) isn't charged
  - SOAP / REST APIs to manage instances
  - Linux AMIs available, other operating systems are being ported
    - But there's always QEMU…

Amazon EC2: Capacity on demand

- Most useful when:
  - You have "variable" demand for a service
  - You have needs which don't require sustained infrastructure
  - You have needs which can be met in a "sandbox"

Amazon EC2: Capacity on demand

- Recent developments:
  - Expanded instance specifications
    - 1, 2, or 4 cores, and a consistent CPU measuring scheme
    - 32 & 64 bit virtual CPUs
    - 1.7, 7.5, or 15 GB of RAM per instance
  - Alternate companies starting to provide similar services
    - With SLAs and without the "beta" label

Infrastructure Differences

- Instances are on dynamic IP addresses sitting outside the local campus network
  - Handshake & heartbeat processes are useful
- Instances have ephemeral storage space
  - Lack of local persistence is not a barrier to entry
  - Be careful of data stored on the ephemeral drive
- Instances need to be monitored for availability
  - Build in safeguards to ensure instances are both available when you need them, and shut down when you don't

Practical challenges

- Applications share platforms without knowing it
  - This can defeat careful performance tuning
  - In fact little is understood about tuning applications to perform well on virtualized platforms like EC2
- Xen itself continues to need work
  - Device drivers run in special dedicated VM partitions
  - Poses a number of strange new security challenges…

11

Practical challenges

- Front end: increasing a "thin client"
  - O/S on your computer focuses on display
  - Applications are running on the hosted VMM in the data center, accessed via the network
- This can mean very slow, creaky graphics performance because traditional graphics interfaces
  - Assume local applications
  - Assume DMA access to bit-mapped images and textures

12

Practical challenges: Security

- Consider a hospital
  - Perhaps doctors, nurses carry a secure id card
  - They put it in a card reader on machine they will use
- With a "real" machine, card has many properties of a TPM…
  - But if the platform has been virtualized and is running at Amazon on EC2, physical security mechanisms can't be used anymore
  - A step backwards relative to TPM-style guarantees

13

Challenges for Amazon itself?

- These revolve around power costs and savings
  - Increasingly clear that the true cost of running a data center is dominated by electric power
  - Leads to a focus on ways to reduce power consumption
- Look at talk by James Hamilton (Microsoft Cloud Computing expert now working at Amazon.com)
  - Talk was delivered at LADIS 2008
  - http://www.cs.cornell.edu/projects/ladis2008/materials/JamesRH_Ladis2008.pdf

14