Today: Data Centers & Cloud Computing

- Data Centers
- Cloud Computing

Data Centers

- Large server and storage farms
  - Used by enterprises to run server applications
  - Used by Internet companies
    - Google, Facebook, Youtube, Amazon…
  - Sizes can vary depending on needs

Data Center Architecture

- Traditional: applications run on physical servers
  - Manual mapping of apps to servers
    - Apps can be distributed
    - Storage may be on a SAN or NAS
  - IT admins deal with "change"
- Modern: virtualized data centers
  - App run inside virtual servers; VM mapped onto physical servers
  - Provides flexiblility in mapping from virtual to physical resources

Virtualized Data Centers

- Resource management is simplified
  - Application can be started from preconfigured VM images / appliances
  - Virtualization layer / hypervisor permits resource allocations to be varied dynamically
  - VMs can be migrated without application down-time

Workload Management

- Internet applications => dynamic workloads
- How much capacity to allocate to an application?
  - Incorrect workload estimate: over- or under-provision capacity
  - Major issue for internet facing applications
    - Workload surges / flash crowds cause overloads
    - Long-term incremental growth (workload doubles every few months for many newly popular apps)
  - Traditional approach: IT admins estimate peak workloads and provision sufficient servers
    - Flash-crowd => react manually by adding capacity
      - Time scale of hours: lost revenue, bad publicity for application

Dynamic Provisioning

- Track workload and dynamically provision capacity
- Monitor -> Predict -> Provision
- Predictive versus reactive provisioning
  - Predictive: predict future workload and provision
  - Reactive: react whenever capacity falls short of demand
- Traditional data centers: bring up a new server
    - Borrow from Free pool or reclaim under-used server
- Virtualized data center: exploit virtualization to speed up application startup time
  - How is this done?

Energy Management in Data Centers

- Energy: major component of operational cost of data centers
  - Large data centers have energy bills of several million $.
  - Where does it come from?
    - Power for servers and cooling
- Data centers also have a large carbon footprint
- How to reduce energy usage?
- Need energy-proportional systems
  - Energy proportionality: energy use proportional to load
  - But: current hardware not energy proportional

Energy Management

- Many approaches possible
- Within a server:
  - Shut-down certain components (cores, disks) when idling or at low loads
  - Use DVFS for CPU
- Most effective: shutdown servers you don't need
  - Consolidate workload onto a smaller # of servers
  - Turn others off
- Thermal management: move workload to cooling or move cooling to where workloads are
  - Requires sensors and intelligent cooling systems

Container-based Data Centers

- Modular design
- No expensive buildings needed
- Plug and play: plug power, network, cooling vent

Example: Container DC

- Courtesy: Dan Reed, Microsoft
  - Talk at NSF workshop
- Benefits of MS Gen 4 data ctr
  - Scalable
  - Plug and play
  - Pre-assembled
  - Rapid deployment
  - Reduced construction

Cloud Computing

- Data centers that rent servers/ storage
- Cloud: virtualized data center with self-service web portal
- Any one with a "credit card" can rent servers
- Automated allocation of servers

- Use virtualized architecture
- Examples: Amazon EC2, Azure, New servers

Cloud Models

- Private clouds versus Public Clouds
  - Who owns and runs the infrastructure?

- What is being rented?
  - Infrastructure as a service (rent barebone servers)
  - Platform as a service (google app engine)
  - Software as a service (gmail, online backup, Salesforce.com)

Pricing and Usage Model

- Fine-grain pricing model
  - Rent resources by the hour or by I/O
  - Pay as you go (pay for only what you use)
- Can vary capacity as needed
  - No need to build your own IT infrastructure for peaks needs

Amazon EC2 Case Study

- Virtualized servers
  - Different sizes / instances
- Storage: Simple storage service (S3)
  - Elastic block service (EBS)
- Many other services
  - Simple DB
  - Database service
  - Virtual private cloud