Cloud Computing and GIS

IT4GIS
Keith T. Weber, GISP
GIS Director
ISU-GIS Training and Research Center

Goal of this presentation

- Describe and demystify "The Cloud"
  - What is it?
  - How is it different from "web services"?
  - ROI and TCO case studies

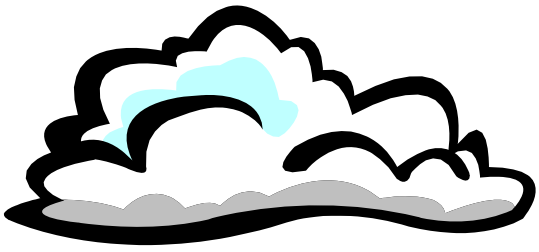
What is *The Cloud*

- Once upon a time…
  - *Sam the server man* discovered his servers were underutilized
  - To optimize their utilization rate, he made his servers available to others
  - This was really nothing new, as ISP's had been around for decades

But…

- Sam's Servers did more than just host HTML pages, they also:
  - Provided infrastructure solutions
  - And hosted services that his customers were not capable of hosting
  - Essentially, they were "servers for rent"
- This commercialization of web service hosting became known as…

Cloud Computing



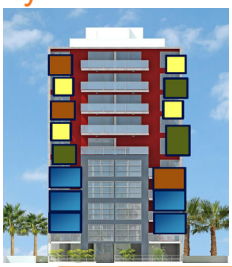
By Definition

- Cloud computing is…
  - On-demand
  - Self-service services
  - Delivered in a metered fashion via a network (i.e., the Internet)
- Cloud computing follows…
  - A multi-tenancy model
  - Within a virtualized, elastic environment

TERMS

Multi-tenancy

- Virtualized servers are used by many

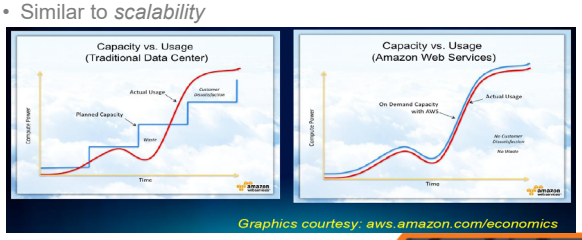


Virtualization

- Similar to logical hard drives, virtualization converts one physical server into many (virtual) servers
- Doing this requires:
  - Physical host server
  - Host OS + Virtualization software (e.g., Hypervisor, Hyper-V)
  - Management suite software

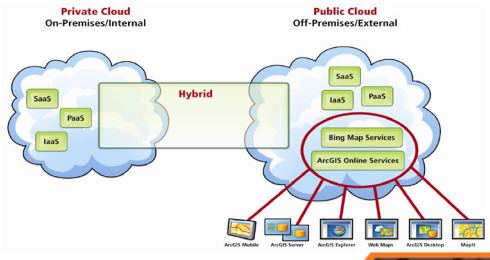
Elasticity

- Similar to *scalability*



Graphics courtesy: aws.amazon.com/economics

Cloud Deployment Models



The Hybrid Model

What *services* are offered in the Cloud?

- If its not html hosting, what services are offered?

aaS

- as a Service

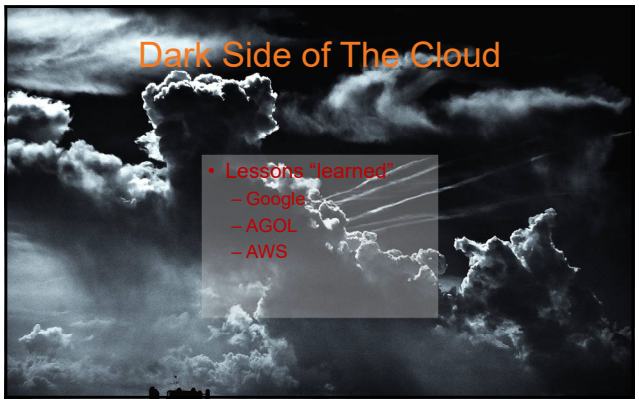
Where does GIS fit aaS?

- Brainstorm…

- **Software as a Service:**
  - ArcLogistics Online
  - Business Analyst Online
  - ArcGIS Explorer Online
- **Platform as a Service**
  - ArcGIS.com / ArcGIS Online
  - ArcGIS Web Mapping
  - ArcGIS Server with Cloud Infrastructure
- **Managed Services**

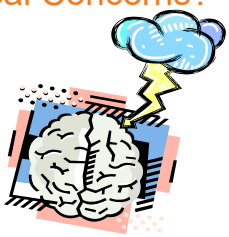
SaaS and the Changing Face of Software

- An increasing number of software applications are hosted in the cloud (SaaS)
- They are offered to the end-user as a subscription
  - They are, of course, dependent upon a good Internet connection to function
- They also put the software developer in charge! Consider this:
  - Compatibility with third party apps/automation scripts
  - Compatibility with Enterprise business constraints



Real Concerns?

- Brainstorm!



To learn more about this watch this TECH talk, Connecting GIS (https://youtu.be/lwQDb2oTzQ0)

How does *The Cloud* differ from Web Services we already learned about?

- **Brainstorm**
- By definition cloud computing is…
  - On-demand
  - Self-service services
  - Delivered in a metered fashion via a network (i.e., the Internet)
- Cloud computing follows…
  - A multi-tenancy model
  - Within a virtualized, elastic environment

Let's Compare

Cloud Servers

- **On-demand**
- **Self-service**
- **Delivered in a metered fashion**
- Multi-tenancy
- Virtualized
- **Elastic**

Virtualized Servers

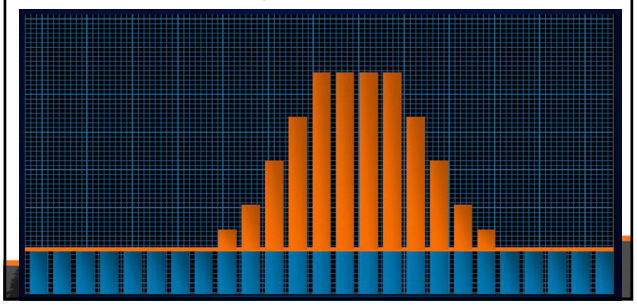
- ~~On-demand~~
- ~~Self-service~~
- ~~Delivered in a metered fashion~~
- Multi-tenancy
- Virtualized
- ~~Elastic~~

ROI and TCO Scenarios

- TCO?
- ROI…
  - Case study #1, a small Idaho county wants to make GIS maps of the county available via the web
  - Case study #2, large research university wants to make GIS maps available via the web

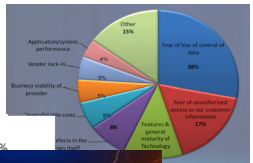
The Partly Cloudy Approach

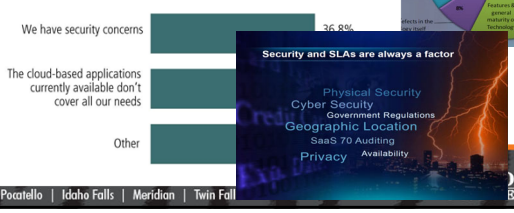
- Own the base, rent the spike



Considerations

- Reasons for **not** using the cloud
  - Security
  - Service Level Agreement





PART TWO

HIGH PERFORMANCE COMPUTING

Goal of this presentation

- Introduce you to an another world of computing, analysis, and opportunity
- Encourage you to learn more!

Some Terminology Up-Front

- Supercomputing
- HPC
- CI

Acknowledgements

- Much of the material presented here, was originally created by Henry Neeman at Oklahoma University and OSCER

What is Supercomputing?

- *Supercomputing* is the biggest, fastest computing today.
- Likewise, a *supercomputer* is one of the biggest, fastest computers today.
- So, the definition of supercomputing is constantly changing.
- Rule of Thumb: A supercomputer is typically 100 X as powerful as a PC.

What is a Supercomputer?

- A cluster of small computers, each called a *node*, hooked together by an *interconnection network* (*interconnect* for short).
- A cluster needs software that allows the nodes to communicate across the interconnect.
- But what a cluster is … is all of these components working together as if they're one big computer ... a super computer.

For example: Dell Intel Xeon Linux Cluster

- 1,076 Intel Xeon CPU chips/4288 cores
- 8,800 GB RAM
- ~130 TB globally accessible disk
- Linux OS
- Peak speed: 34.5 TFLOPs*
  - *TFLOPs: trillion floating point operations (calculations) per second

sooner.oscer.ou.edu

What is Supercomputing About?

Size

Speed

Laptop

Size…

- Many problems that are interesting to scientists **can't fit on a PC**
  - usually because they need more than 32 GB of RAM, or more than a few TB of disk.

Speed…

- Many problems that are interesting to scientists and engineers would take a long time to run on a PC.
  - Days, weeks and even months.
  - But a problem that would take 1 month on a PC might take only a few hours on a supercomputer
- **How can a supercomputer do this?**

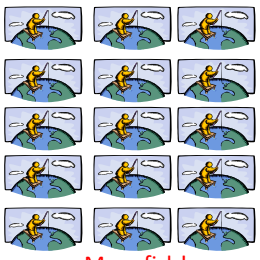
How a cluster works together:

PARALLELISM

Parallelism

Parallelism means doing multiple things at the same time

Less fish …

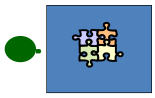


More fish!

Understanding Parallel Processing

THE JIGSAW PUZZLE ANALOGY

Serial Computing

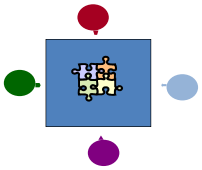


- We are very accustom to serial processing. It can be compared to building a jigsaw puzzle by yourself.
- Suppose you want to complete a jigsaw puzzle that has 1,000 pieces.
- This will take a certain amount of time…let's just say, one hour

Shared Memory Parallelism

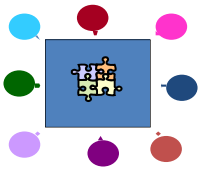
- If Scott sits across the table from you, then he can work on half the puzzle and you can work on your half.
- Once in a while, you'll both reach into the pile of pieces at the same time (you'll ***contend*** for the same resource), which will cause you both to slowdown.
- And from time to time you'll have to work together (***communicate***) at the interface between his half and yours. The speedup will be *nearly* 2-to-1: **Together it will take about 35 minutes instead of 30.**

The More the Merrier?



- Now let's put Paul and Charlie on the other two sides of the table.
- Each of you can work on a part of the puzzle, but there'll be a lot more **contention** for the shared resource (the pile of puzzle pieces) and a lot more **communication** at the interfaces.
- You will achieve noticeably less than a 4-to-1 increase in speed.
- But you'll still have an improvement, perhaps 20 minutes instead of 1 hour.

Diminishing Returns



- If we now put Dave, Tom, Horst, and Ella at the corners of the table, there's going to be a much more **contention** for the shared resource, and a lot of **communication** at the many interfaces.
- The speedup will be much less than we'd like; you'll be lucky to get 5-to-1.
- Adding more and more workers to a shared resource is eventually going to have a diminishing return.

Distributed Parallelism

- Let's try something a little different.
  - Set up two tables
  - You will sit at one table and I will set at the other.
  - We will put half of the puzzle pieces on your table and the other half of the pieces on my end of the table
- Now you can work completely independently, without any **contention** for a shared resource.
- **BUT**, the cost of communication is **MUCH** higher, and you need the ability to split up (***decompose***) the puzzle pieces correctly.

More Distributed Processors

- It's easy to add more processors (these are the people assembling the puzzle) in distributed parallelism.
- But you must be aware of the need to:
  - **decompose** the problem and
  - Allow for **communication** among/between the processors.
- Also, as you add more processors, it may be harder to ***load balance*** the amount of work that each processor gets.

Why Parallelism Is Good

- As the number of processing units working on a problem grows, we can solve:
  - The same problem in less time (think of a tree)
  - Bigger problems (think of a forest)

GIS Applications?

- Can you think of a GIS workflow that could be parallelized easily?
  - Lidar and LAS Tools

Interested? Curious?

- To learn more, or to get involved with supercomputing there is a host of opportunities awaiting you
  - Get to know ISU's Campus Champions
  - Ask about internships at INL C3
  - Learn C (not C++, but C) or Fortran
  - Python is coming along too!
  - Learn UNIX (Linux)

Questions?

Time for our final 2-minute Write!