# High performance computing, cyberinfrastructure and GIS

Dr. Junjun Yin

*CyberGIS Center for Advanced Digital and Spatial Studies*
*Department of Geography and Geographic Information Science*
*National Center for Supercomputing Applications*
*University of Illinois at Urbana-Champaign, IL, 61801, USA*
*jyn@illinois.edu*

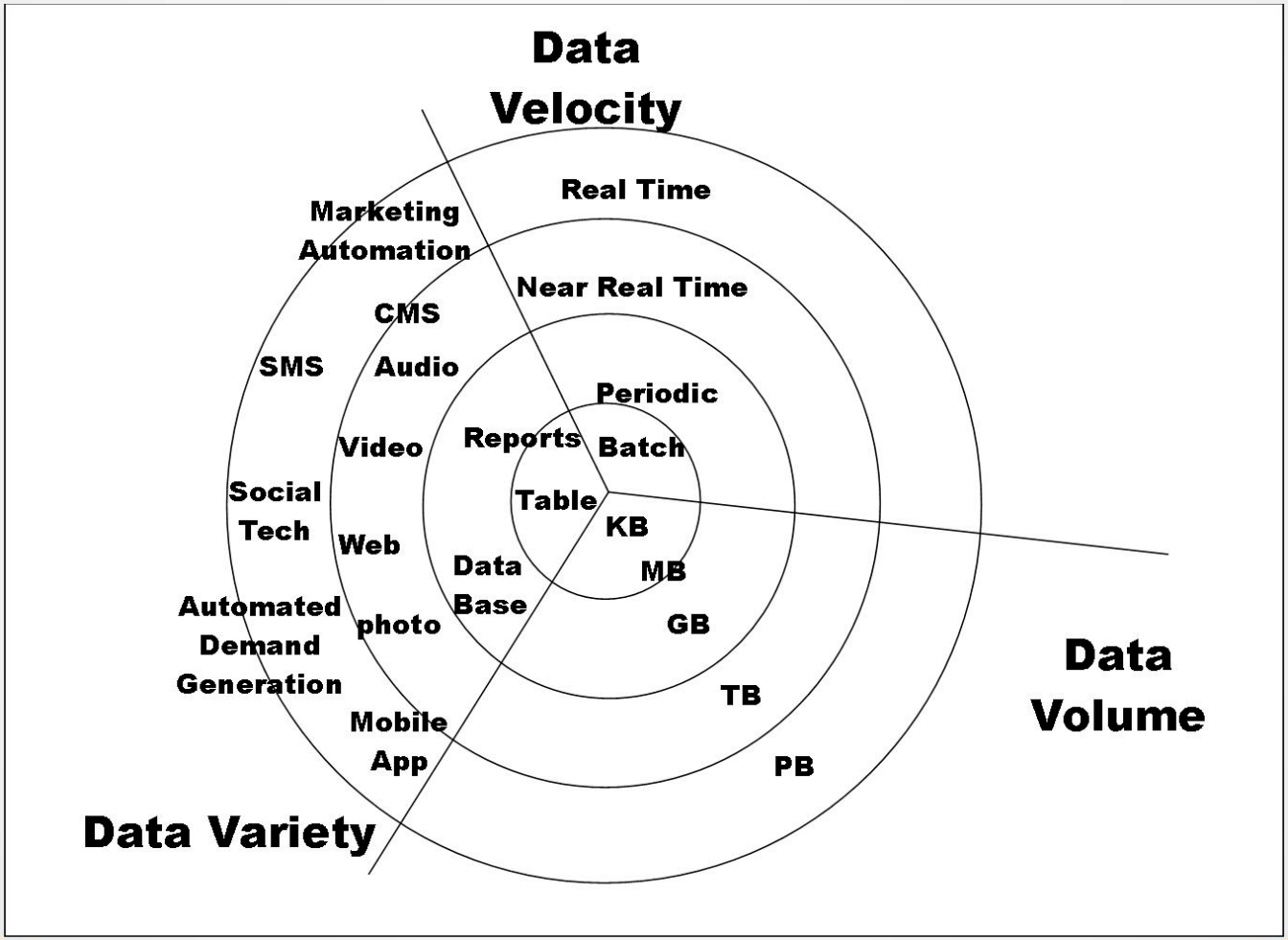
*January 26, 2016*

# Outline

- High Performance Computing (HPC)
  - "High Performance Computing most generally refers to the practice of aggregating computing power in a way that delivers much higher performance than one could get out of a typical desktop computer or workstation in order to solve large problems in science, engineering, or business." [1]
- cyberinfrastructure
  - The available HPC resources regarding computing platforms and frameworks
  - ROGER
- GIS
  - Leverage cyberinfrastructure to address challenges in geospatial Big Data
  - Utilize geo-processing as services
  - Tailor cyberinfrastructure resources for specific geospatial problem solving, e.g., choosing right computing platform and paradigm

1. http://insidehpc.com/hpc-basic-training/what-is-hpc/

# Need for HPC



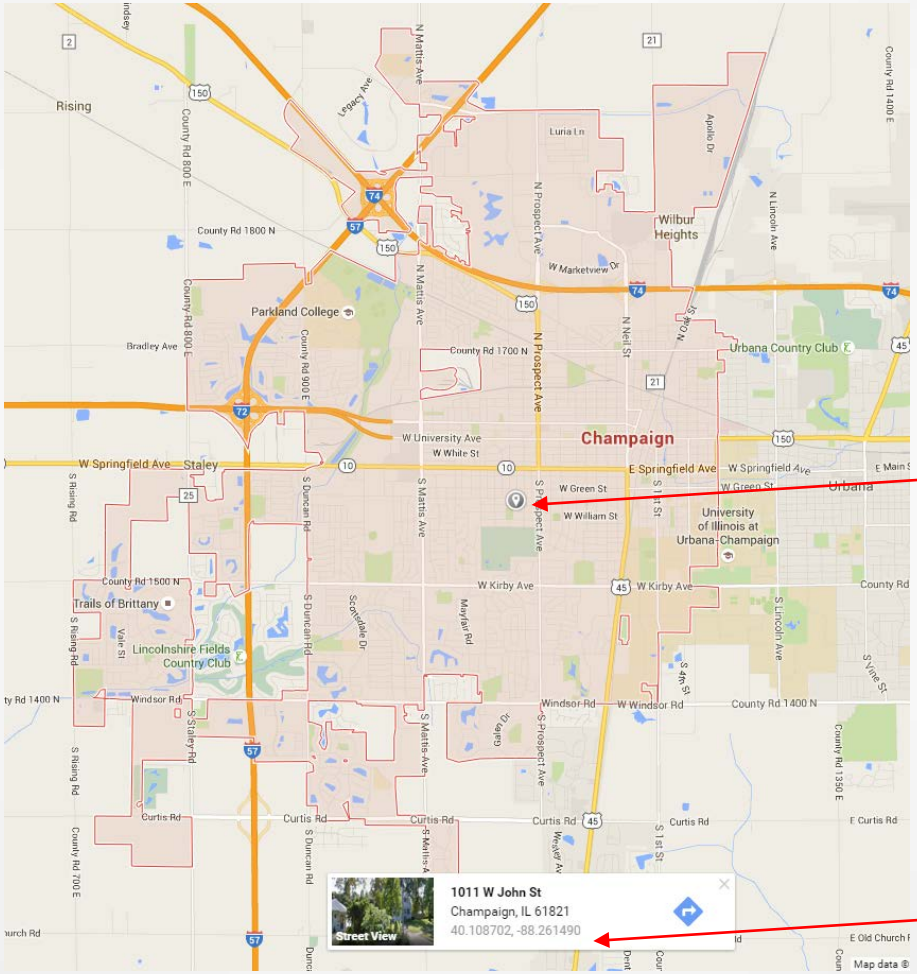

The 3Vs of Big Data

Source: http://velvetchainsaw.com/2012/07/20/three-vs-of-big-data-as-applied-conferences/

# Simple Geospatial Scenarios

- Geocoding and reverse-geocoding
  - Geocoding: address to geographical coordinates
  - Reverse-geocoding: geographical coordinates to address
  - Example database: GeoNames (http://www.geonames.org/)
- Navigation
  - Shortest/fastest path between origin and destination
- Geoprocessing
  - GIS operations to manipulate spatial data, a typical example is the ArcGIS geoprocessing toolbox
- Geospatial Data Integration
  - Integrating multiple geographical data layers for spatial analysis (e.g., multi-criteria analysis)

# Simple Geospatial Scenarios




An example of geocoding using Google Maps API

# Geospatial Big Data: New York Taxi Data

January 2013 Taxi Trips from FOIA (Freedom of Information Act) request from Chris Whong

(http://www.andresmh.com/nyctaxitrips/)

**Trip:**

```
medallion, hack_license, vendor_id, rate_code,
store_and_fwd_flag, pickup_datetime,
dropoff_datetime, passenger_count, trip_time_in_secs,
trip_distance, pickup_longitude, pickup_latitude,
dropoff_longitude, dropoff_latitude
```

**Fare:**

```
medallion, hack_license, vendor_id, pickup_datetime,
payment_type, fare_amount,surcharge, mta_tax,
tip_amount, tolls_amount, total_amount
```

# What does data look like?

jun@jun-ncsa: ~/Desktop

```
DFD2202EE08F7A8DC9A57B02ACB81FE2,51EE87E3205C985EF8431D850C786310,CMT,1,N,2013-01-07 23:54:15,2013-01-07 23:58:20,2,244,.70,-73.974602,40.759945,-73.984734,40.759388
DFD2202EE08F7A8DC9A57B02ACB81FE2,51EE87E3205C985EF8431D850C786310,CMT,1,N,2013-01-07 23:25:03,2013-01-07 23:34:24,1,560,2.10,-73.97625,40.748528,-74.002586,40.747868
20D9ECB2CA0767CF7A01564DF2844A3E,598CCE5B9C1918568DEE71F43CF26CD2,CMT,1,N,2013-01-07 15:27:48,2013-01-07 15:38:37,1,648,1.70,-73.966743,40.764252,-73.983322,40.743763
496644932DF3932605C22C7926FF0FE0,513189AD756FF14FE670D10B92FAF04C,CMT,1,N,2013-01-08 11:01:15,2013-01-08 11:08:14,1,418,.80,-73.995804,40.743977,-74.007416,40.744343
0B57B9633A2FECD3D3B1944AFC7471CF,CCD4367B417ED6634D986F573A552A62,CMT,1,N,2013-01-07 12:39:18,2013-01-07 13:10:56,3,1898,10.70,-73.989937,40.756775,-73.86525,40.77063
2C0E91FF20A856C891483ED63589F982,1DA2F6543A62B8ED934771661A9D2FA0,CMT,1,N,2013-01-07 18:15:47,2013-01-07 18:20:47,1,299,.80,-73.980072,40.743137,-73.982712,40.735336
2D4B95E2FA7B2E85118EC5CA4570FA58,CD2F522EEE1FF5F5A8D8B679E23576B3,CMT,1,N,2013-01-07 15:33:28,2013-01-07 15:49:26,2,957,2.50,-73.977936,40.786983,-73.952919,40.80637
E12F6AF991172EAC3553144A0AF75A19,06918214E951FA0003D1CC54955C2AB0,CMT,1,N,2013-01-08 13:11:52,2013-01-08 13:19:50,1,477,1.30,-73.982452,40.773167,-73.964134,40.773815
E12F6AF991172EAC3553144A0AF75A19,06918214E951FA0003D1CC54955C2AB0,CMT,1,N,2013-01-08 09:50:05,2013-01-08 10:02:54,1,768,.70,-73.99556,40.749294,-73.988686,40.759052
78FFD9CD0CDA541F335EF8B38FB494D6,E949C583ECF62C8F03FDCE1484954A08,CMT,1,N,2013-01-10 12:07:08,2013-01-10 12:17:29,1,620,2.30,-73.971497,40.791321,-73.964478,40.775921
237F49C3ECC11F5024B254268F054384,93C363DDF8ED9385D65FAD07CE3F5F07,CMT,1,N,2013-01-07 07:35:47,2013-01-07 07:46:00,1,612,2.30,-73.98851,40.774307,-73.981094,40.755325
3349F919AA8AE5DC9C50A3773EA45BD8,7CE849FEF67514F080AF80D990F7EF7F,CMT,1,N,2013-01-10 15:42:29,2013-01-10 16:04:02,1,1293,3.20,-73.994911,40.723221,-73.971558,40.761612
3349F919AA8AE5DC9C50A3773EA45BD8,7CE849FEF67514F080AF80D990F7EF7F,CMT,1,N,2013-01-10 14:27:28,2013-01-10 14:45:21,1,1073,4.40,-74.010391,40.708702,-73.987846,40.756104
4C005EEBAA7BF26B84B21586332488A2,351BE7D984BE17DB2FA80A748E816472,CMT,1,N,2013-01-07 22:09:59,2013-01-07 22:19:50,1,591,1.70,-73.973732,40.756287,-73.998413,40.756832
7D99C30FCE69B1A9DD27E2AEAC9BFA0C,460C3F57DD9CB2265DB75B14CD70224D,CMT,1,N,2013-01-07 17:18:16,2013-01-07 17:20:55,1,158,.70,-73.968925,40.767704,-73.96199,40.776566
E6FBF80668FE0611AEA44FD9574A7E32,36773E80775F26CD1158EB5450A61C79,CMT,1,N,2013-01-07 06:08:51,2013-01-07 06:13:14,1,262,1.70,-73.96212,40.769737,-73.979561,40.75539
0C5296F3C8B16E702F8F2E06F5106552,D2363240A9295EF570FC6069BC4F4C92,CMT,1,N,2013-01-07 22:25:46,2013-01-07 22:36:56,1,669,2.30,-73.989708,40.756714,-73.977615,40.787575
D8AADA4E722C87C10E6096546126 30DD,8F1E4F779A4D67431C3FCFA4E70A87DC,CMT,2,N,2013-01-10 23:41:51,2013-01-11 00:09:11,1,1640,17.50,-73.78331,40.648766,-73.988914,40.748207
B6FC8357E0D53B0F1FA897D536A20F5C,113A5BA8513934DEE97A342E3535DE96,CMT,1,N,2013-01-07 18:05:36,2013-01-07 18:23:50,4,1094,3.80,-73.956505,40.771278,-73.996368,40.73246
E12F6AF991172EAC3553144A0AF75A19,06918214E951FA0003D1CC54955C2AB0,CMT,1,N,2013-01-08 13:29:25,2013-01-08 13:37:52,1,507,1.20,-73.95578,40.77932,-73.967285,40.763344
4C005EEBAA7BF26B84B21586332488A2,351BE7D984BE17DB2FA80A748E816472,CMT,1,N,2013-01-07 21:13:02,2013-01-07 21:22:31,1,568,1.10,-73.978439,40.764679,-73.977684,40.777004
E9BA494DC1A1F26D186394EFBB8327CF,BBF604B3A2426BFBAAC982AAAA0BF1B5,CMT,1,N,2013-01-01 18:36:53,2013-01-01 18:39:04,1,131,.50,-73.992172,40.749954,-73.99675,40.744553
E6FBF80668FE0611AEA44FD9574A7E32,36773E80775F26CD1158EB5450A61C79,CMT,1,Y,2013-01-07 08:17:06,2013-01-07 08:22:46,1,340,1.50,-73.992111,40.689701,-74.007156,40.679295
0C5296F3C8B16E702F8F2E06F5106552,D2363240A9295EF570FC6069BC4F4C92,CMT,1,N,2013-01-07 22:39:03,2013-01-07 22:44:16,1,312,1.50,-73.978653,40.787735,-73.96579,40.805321
A3B17384165197E18CA0A1BB61277EE9,B8396B62883EA332EFD2771A6B031D05,CMT,1,N,2013-01-07 06:26:32,2013-01-07 06:28:24,2,111,.50,-73.965317,40.769375,-73.967133,40.763699
8E189DABE265CC03FEF4BFC695B32559,906939B31C0B93768242A58A682412B8,CMT,1,N,2013-01-01 14:02:01,2013-01-01 14:27:39,2,1537,10.20,-73.862709,40.769142,-73.982079,40.762295
312E0CB058D7FC1A6494EDB66D360CD2,7B5156F38990963332B33298C8BAE25E,CMT,1,N,2013-01-05 11:54:49,2013-01-05 12:03:48,1,539,.80,-73.977127,40.74831,-73.990913,40.751053
F1EF8290A54338B1396DD98E38E09143,0FDDEFAFF6FC38BD632B6DDED47DA618,CMT,1,N,2013-01-05 08:16:58,2013-01-05 08:30:20,1,801,6.50,-73.976318,40.682724,-73.917915,40.742664
0F9E0728AB1E40D5CEB0C6EDBF805CCB,8434E8A33D8C0150573FAA00B8A9ABF5,CMT,1,N,2013-01-05 19:04:43,2013-01-05 19:13:58,1,555,2.80,-73.966682,40.761139,-73.938515,40.792332
33A0B414EB87DB2538CB5929BF1EEE0D,8105CBF11B1747C525FA81334375EF5B,CMT,1,N,2013-01-05 17:34:11,2013-01-05 17:48:54,1,882,5.20,-73.948334,40.776482,-73.99176,40.733841
24B56A4A0AC119529DBA559181C14FE4,CD5A75F5F950B3E26D561732332C8A5D,CMT,1,N,2013-01-05 18:59:29,2013-01-05 19:12:21,1,771,2.40,-73.993202,40.724407,-73.984833,40.748741
7E3256C342CAFB3C23D3ABB98B9E3FB6,4E7873C913735088E9BEE85C1D054D31,CMT,1,N,2013-01-05 03:00:46,2013-01-05 03:23:28,1,1361,5.50,-73.9655,40.711113,-73.918327,40.758671
```

# Geospatial Big Data: New York Taxi Data

- Each file contains chunks of data in csv format, ranging from ~1.5 to ~2.5 GB in size
- Each file has about 14 million rows
- What would you do?
  - If we want to know the location information for each pick up geographical coordinates?
  - If you want to determine the busiest road/street when taxi passes by
  - Where is the most dense pick-up/drop-off areas?
  - …

# Computing Platforms

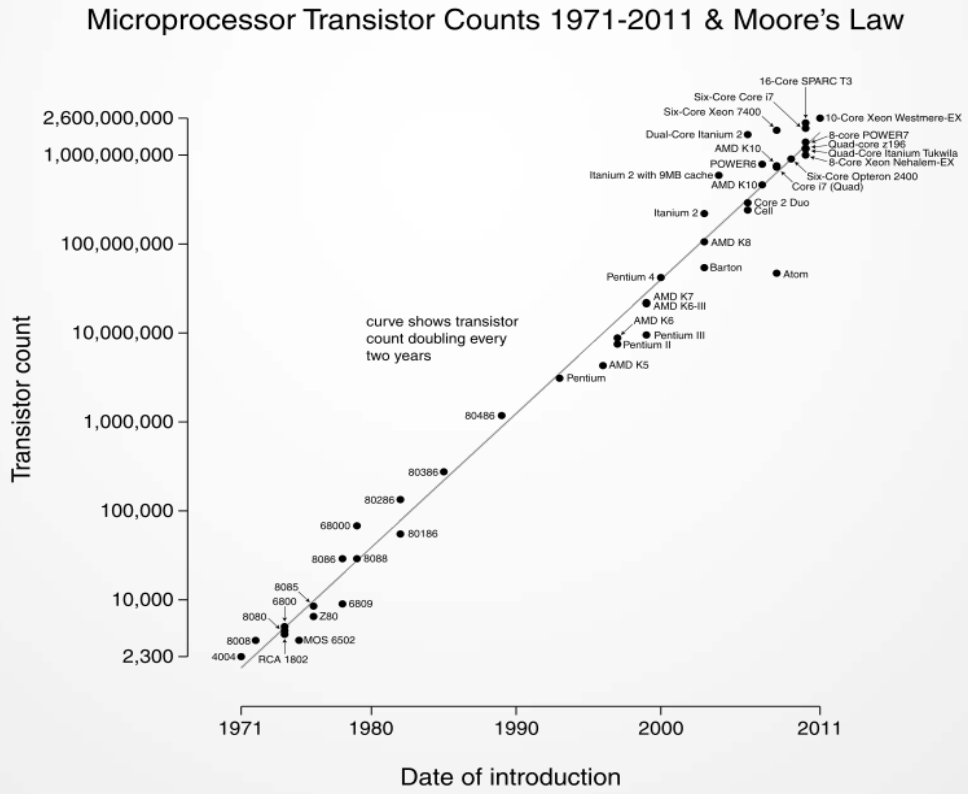
- High Performance computing
  - Performance vs. scalability
  - Mainly realized via parallel computing
- High-throughput computing
  - use of many computing resources over long periods of time to accomplish a computational task
  - Robustness and reliability of jobs over a long-time scale
- Distributed computing
  - Grid computing
  - Cloud computing
  - Distributed databases

# High Performance Computing

- High Performance Computing (HPC) is computation at the cutting-edge of modern computing technology, often done on a supercomputer.
- A supercomputer is in the class of machines that rank among the fastest in the world
  - Rule of thumb: a supercomputer could be defined to be at least 100 times as powerful as a ordinary PC
- Computer performance is measured in FLoating point Operations Per Second (FLOPS or flop/s)
  - $FLOPS = cores \times clock$ rate ×Float point operation/C$ycle$
  - Most microprocessors today can do 4 float point operation per clock cycle. Therefore a 2.5-GHz processor has a theoretical performance of 10 billion FLOPs = 10 GFLOPs.

# Development of supercomputers

- Moore's Law
  - over the history of computing hardware, the number of transistors in a dense integrated circuit doubles approximately every two years.




Source: https://en.wikipedia.org/wiki/Moore%27s_law

# Development of supercomputers

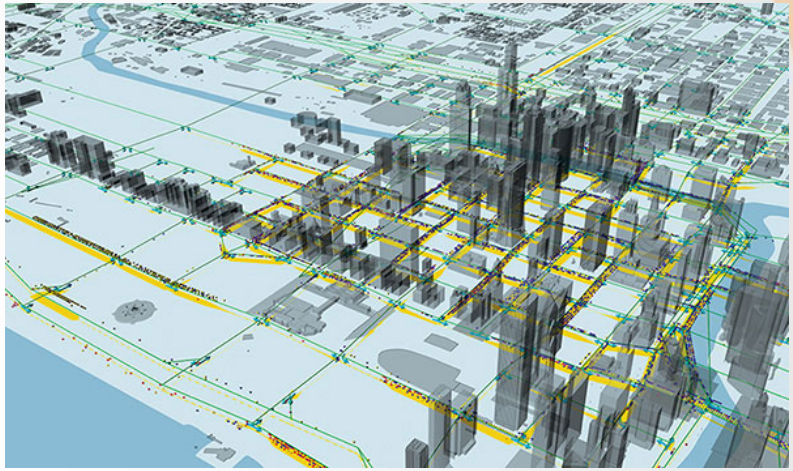
- Your cell phone today is equal to a supercomputer 15 years ago.
  - A8 CPU of Iphone 6/6 plus is around 112 GFLOPS, which can be ranked in Top 200 supercomputers in 2000

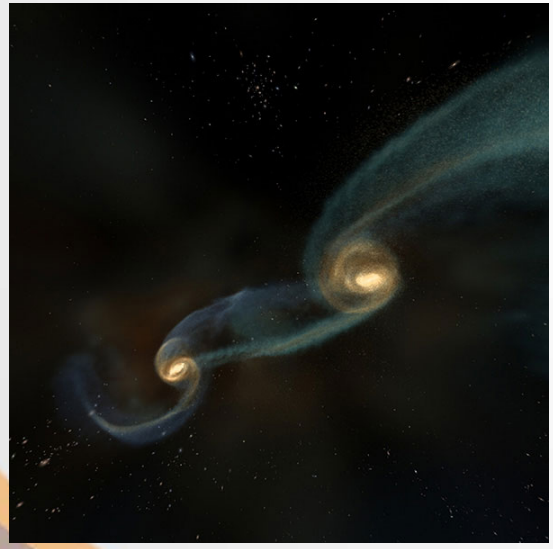

IBM SP power3 375 MHZ (15 years ago)

Iphone 6/6+ today

# HPC Applications

- Simulation of physical phenomena
  - Storm surge prediction
  - Black holes colliding
  - Molecular dynamics





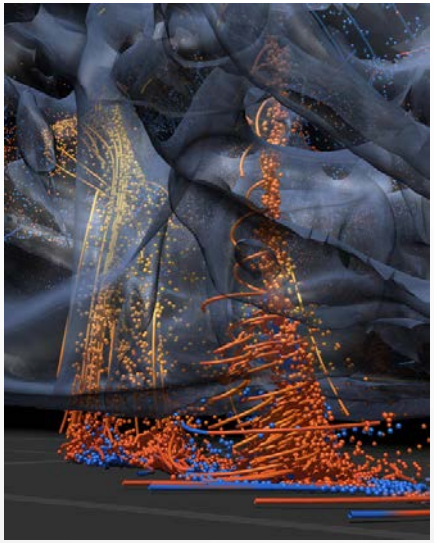


Image source from: http://ncsa.illinois.edu

# Parallel Computing

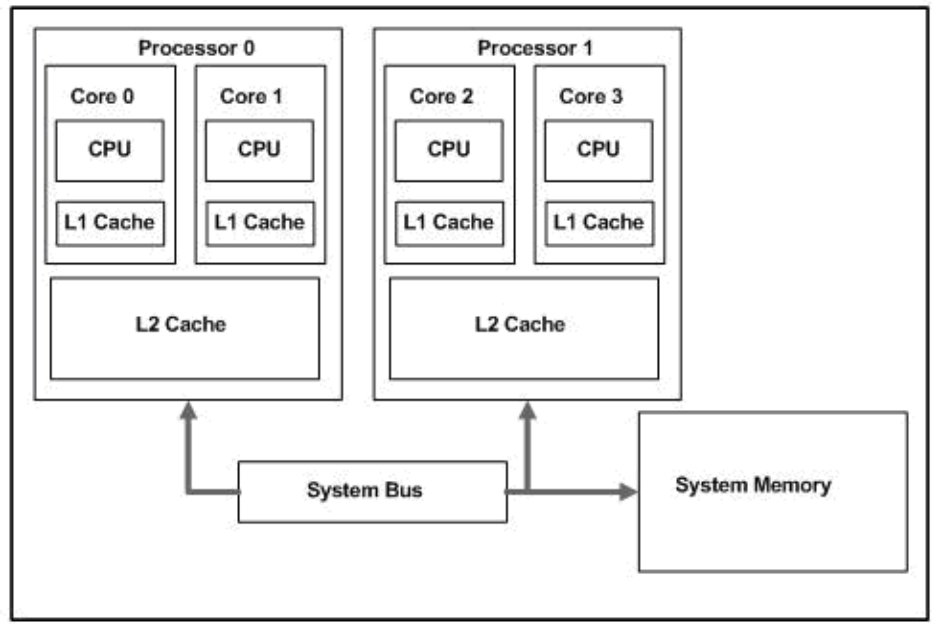
- Parallel computing
  - Multiple processing units (workers) work simultaneously to process a workload
  - The processing units can be tightly or loosely coupled
  - In most cases, the processing units needs to communicate (sharing data) among each other for coordination.
  - Most supercomputers designed for parallel computing

# Why Parallel

- Natural phenomena operate in parallel fashion
- Overcome the bottleneck of CPU speed and bandwidth
- Solve large and complex problems (in many cases cannot be solved by serial computing)
- Take advantage of modern parallel hardware
- Save time/money
- **Limitations**
- Increased complexity of program (difficult debug)
- Communication overhead among processes
- Not all programs can be parallelized

# Parallel Computing Platform

- **Multiprocessor/multicore:**
  - several processors/cores work on data stored in shared memory. (e.g., Intel Core i7 quad-core processor)




Shared Cache Example: A dual-core, dual-processor system

# Parallel Computers

- **Co-processor:**
  - a general-purpose processor supplement the functions of the primary processor (e.g. GPU)



NVIDIA Tesla K20 GPU

- **Hybrid:**
  - Cluster of multicore/multiprocessor nodes with GPUs

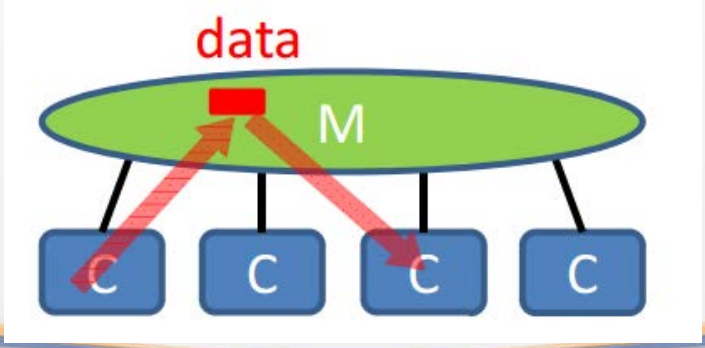
# Parallel Programming Models

There are different parallel programming models according to how data are shared and exchanged among processors.

- Multi-Threading (shared memory)
- Message passing (distributed memory)
- Hybrid

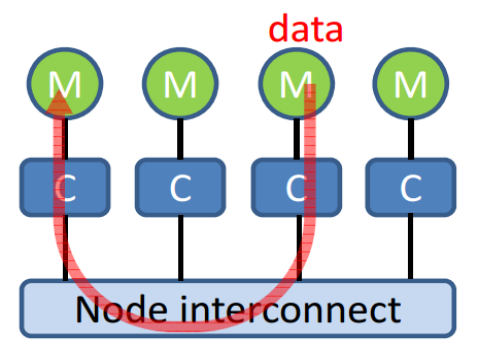
# Multi-threading

- Shared memory
  - All threads can access the global address space
  - Data sharing achieved via writing to/reading from the same memory location.
  - Advantage: no need to explicitly specify the data communication (simpler programming, easy translation from serial code)
  - Disadvantage: be careful for synchronization conflicts, only applies to Symmetric Multi-Processing (SMP) machines
  - Examples:
    - OpenMP (C, C++, Fortran)
    - Multiprocessing modules in Python



# Message Passing

- Distributed memory
  - A set of computing nodes use their own local memory during computation.
  - Data transfer requires specifying the sender, receiver and mode of transfer.
  - Advantage: High flexibility->higher performance
  - Disadvantage: Complex programming, not easy to translate from serial code. Communication overhead.
  - Example
    - Message Passing Interface (MPI) supported by most programming languages

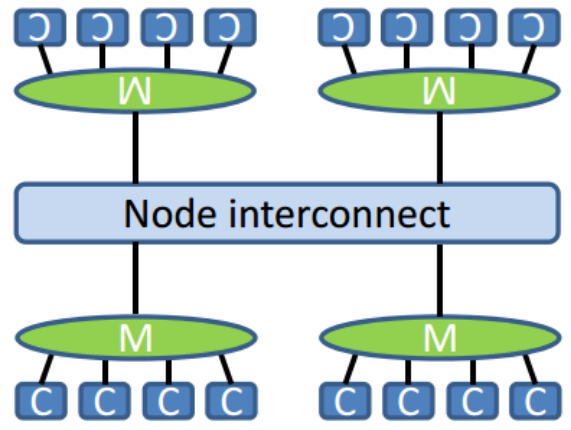


# Hybrid model

- Clusters of SMP (symmetric multi-processing) nodes dominate nowadays
- Two-Level Architecture:

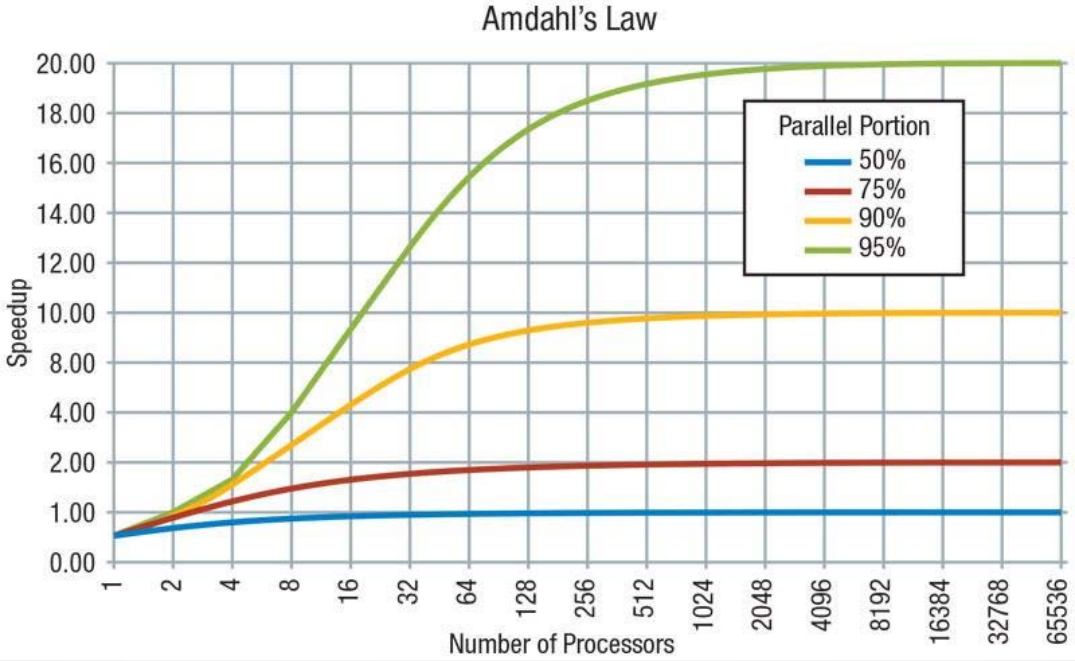
  1: Cores/processors share local memory

  2: Nodes with distributed memory communicate through message passing
- Combining different models, e.g. OpenMP and MPI



# Scalability of Parallel Computing

- T(n,p) is the time to solve a problem of size n using p processors
  - Speedup: S(n,p) = T(n,1)/T(n,p)
  - Efficiency: E(n,p) = S(n,p)/p
  - Scaled Efficiency: SE(n,p) = T(n,1)/T(pn,p)
- Amdahl's Law: Maximum speedup = 1/(1-Parallel portion)

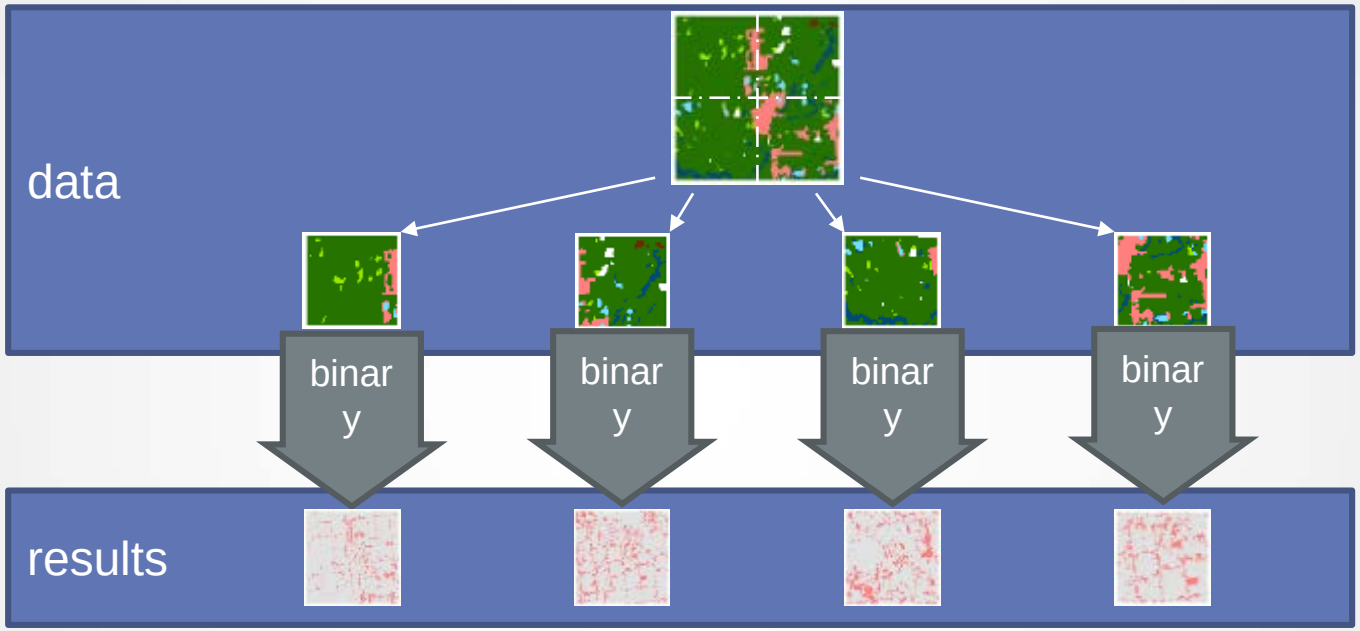


# Data Parallelism

- Data parallelism
  - The same task is run on different data in parallel

- Example: convert all pixels in a raster to binary {1: developed, 0: other land cover}

# Data Parallelism



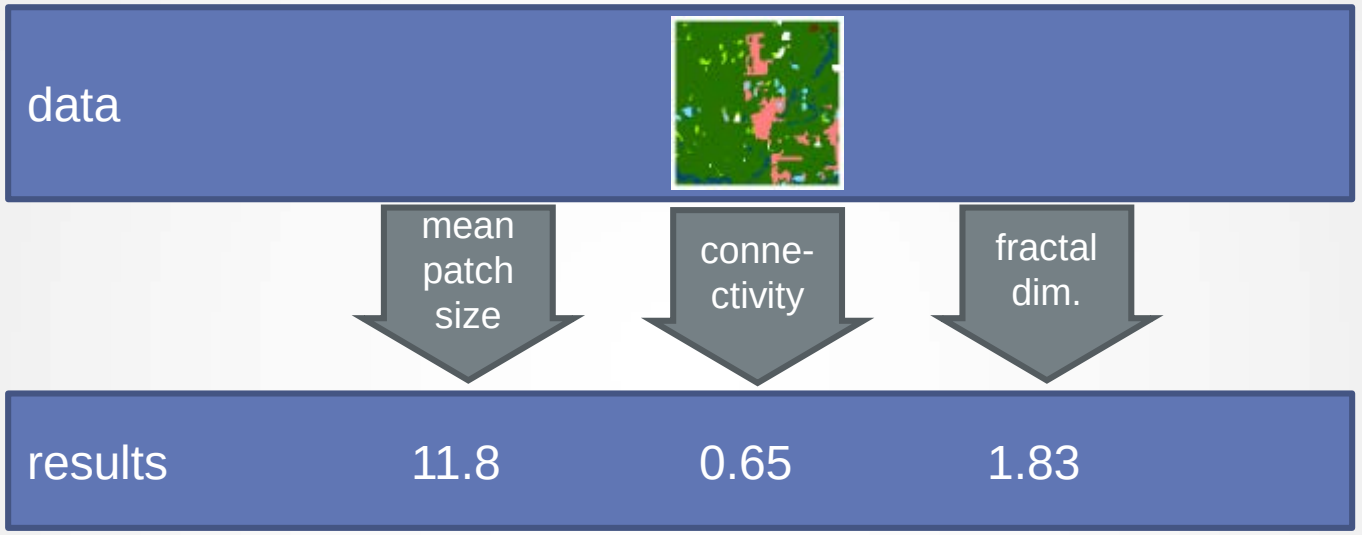

- Can divide the raster into smaller blocks and run the tasks (i.e. binarize each block) in parallel
- No dependencies between the tasks

# Task Parallelism

- Task parallelism
  - Different tasks are run on the same data concurrently

- Example: summary indices of the same raster map, e.g. mean patch size, landscape connectivity, fractal dimension

# Task Parallelism




- Several functions on the same data (especially when each function needs access to the full dataset)
- No dependencies between the tasks

# Hybrid Data-Task Parallelism

- Structured
  - A parallel pipeline of tasks, each of which might be data parallel
- Unstructured
  - Ad hoc combination of threads with no obvious top-level structure

# Distributed Computing

- Distributed data sources
- Collaborative analysis
- Parallel computing
- Cloud computing

# Node

- Discrete unit of computer system, runs its own instance of OS (Ex. Laptop is one node)
- Computing units → Cores: can process separate streams of instructions
- Number of cores in a node = (number of processing chips) x (number of cores on a chip)
- Note: Modern nodes contain processor chips sharing memory, disk

# Cluster

- Collection of machines (nodes) that function in some way as a single resource (e.g. Stampede)
- Nodes of cluster are assigned by a “scheduler”
- Job: Assignment of nodes for a certain user, certain amount of time

# Grid

- Software stack facilitating shared resources across networks, institutions;
- Deals with heterogeneous clusters
- Most grids are cross-institutional groups of clusters, common software, common user authentication

# Grid Computing

- **Grid computing:** combines the different computing resources derived from distributed locations to support a specific application
- **Grid:** generates the connection between different computing resources

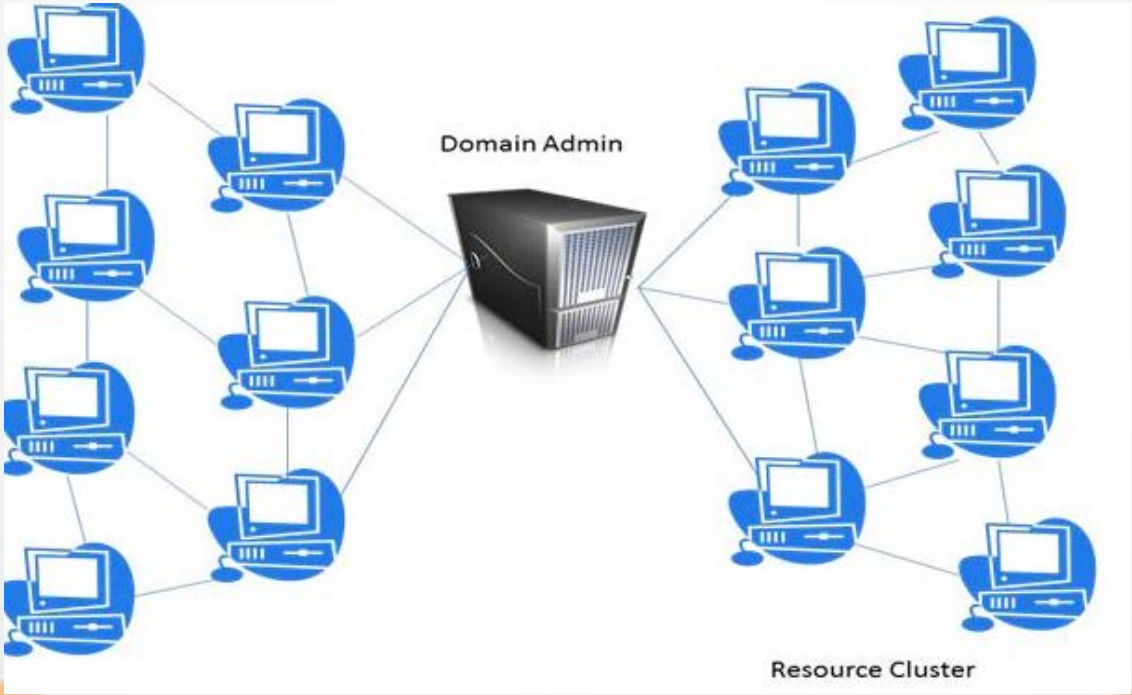
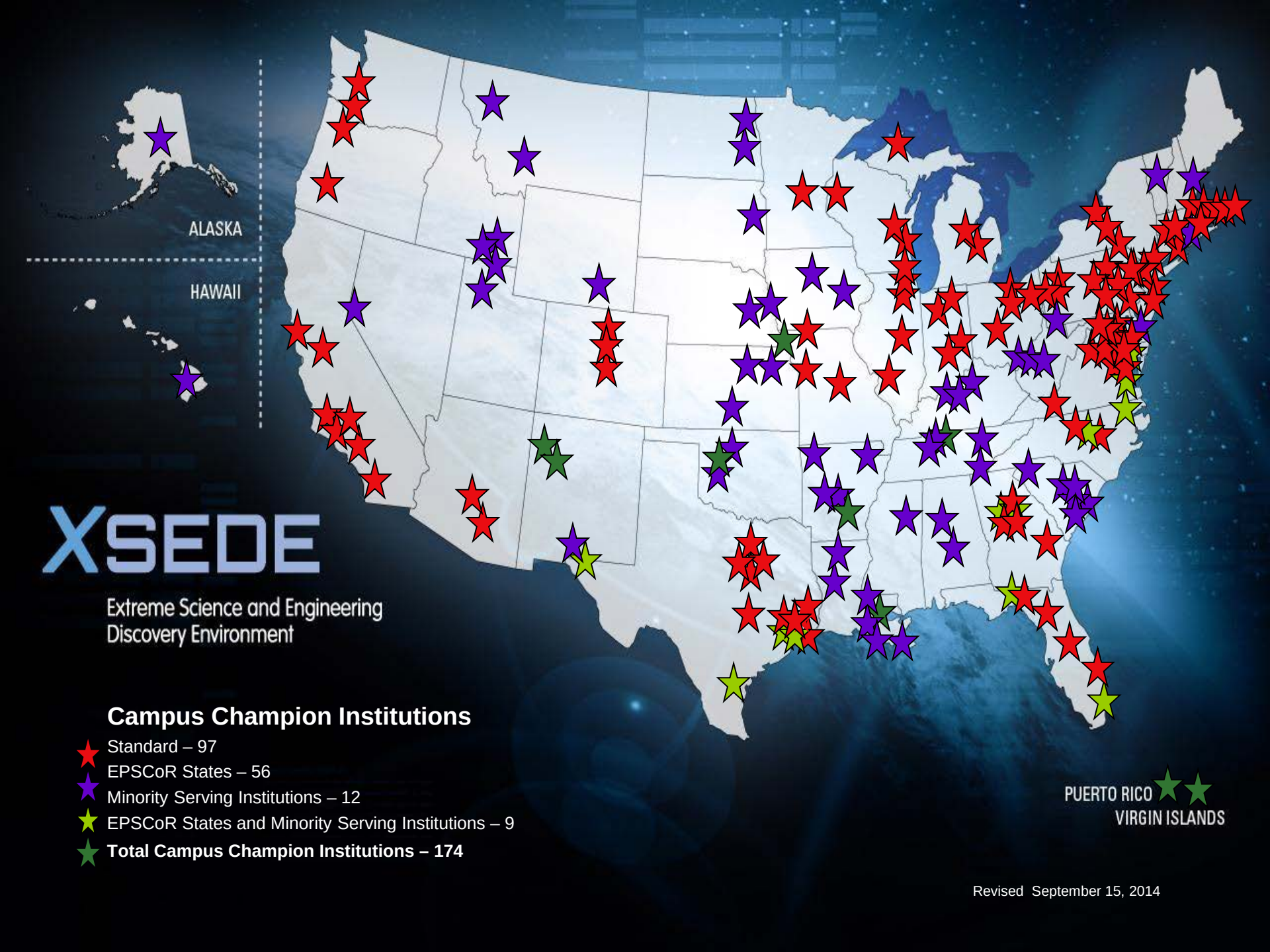



Image source: http://cloudservice24h.com/cloud-and-grid-computing-whats-different/

- Major grids in the U.S.
  - OSG: Institutions organize themselves into virtual organizations (VOs) with similar computing interests. They all install OSG software
  - XSEDE: Similar to OSG, limits usage to dedicated high-performance network


ALASKA
HAWAII
XSEDE
Extreme Science and Engineering Discovery Environment
Campus Champion Institutions
Standard – 97
EPSCoR States – 56
Minority Serving Institutions – 12
EPSCoR States and Minority Serving Institutions – 9
Total Campus Champion Institutions – 174
PUERTO RICO
VIRGIN ISLANDS
Revised September 15, 2014

# XSEDE

- Extreme Science and Engineering Discovery Environment (XSEDE)
- “most advanced, powerful, and robust collection of integrated advanced digital resources and services in the world”
- Five-year, $121-million project is supported by the National Science Foundation
- Originally TeraGrid until July 2011.
- XSEDE is an extension and expansion of this original program.
- 16 supercomputers
- Support short-term and long-term projects
- No monetary cost to scientists
- The goal is to give scientists and researchers the tools for creating scientific discoveries via high-performance computing technologies

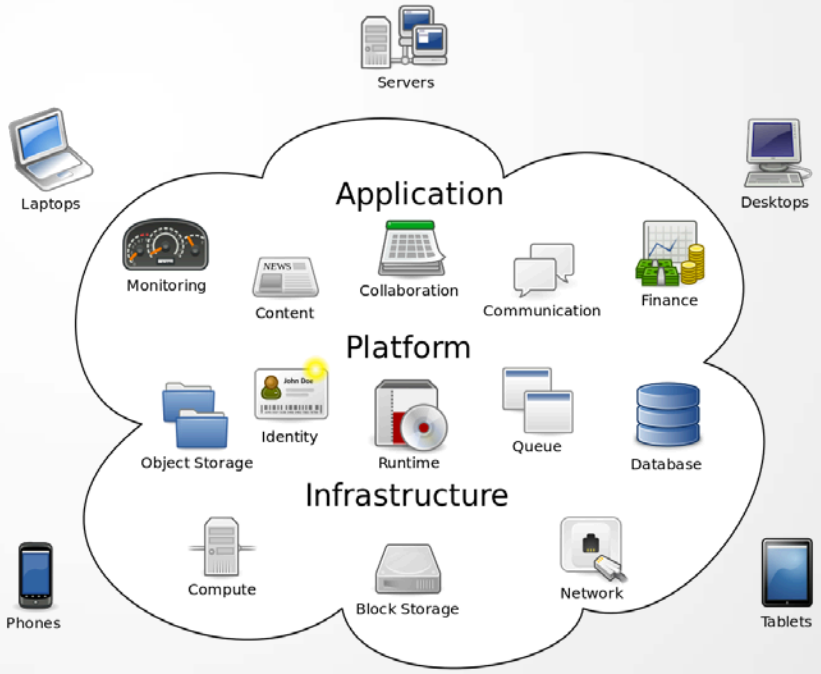
# XSEDE Resources

Four types:

- Computing resources
- Visualization resources
- Storage resources
- High-throughput resources

# Cloud Computing

- Cloud computing
  - Cloud computing is computing in which large groups of remote servers are networked to allow the centralized data storage, and online access to computer services or resources.



# Architectural Model

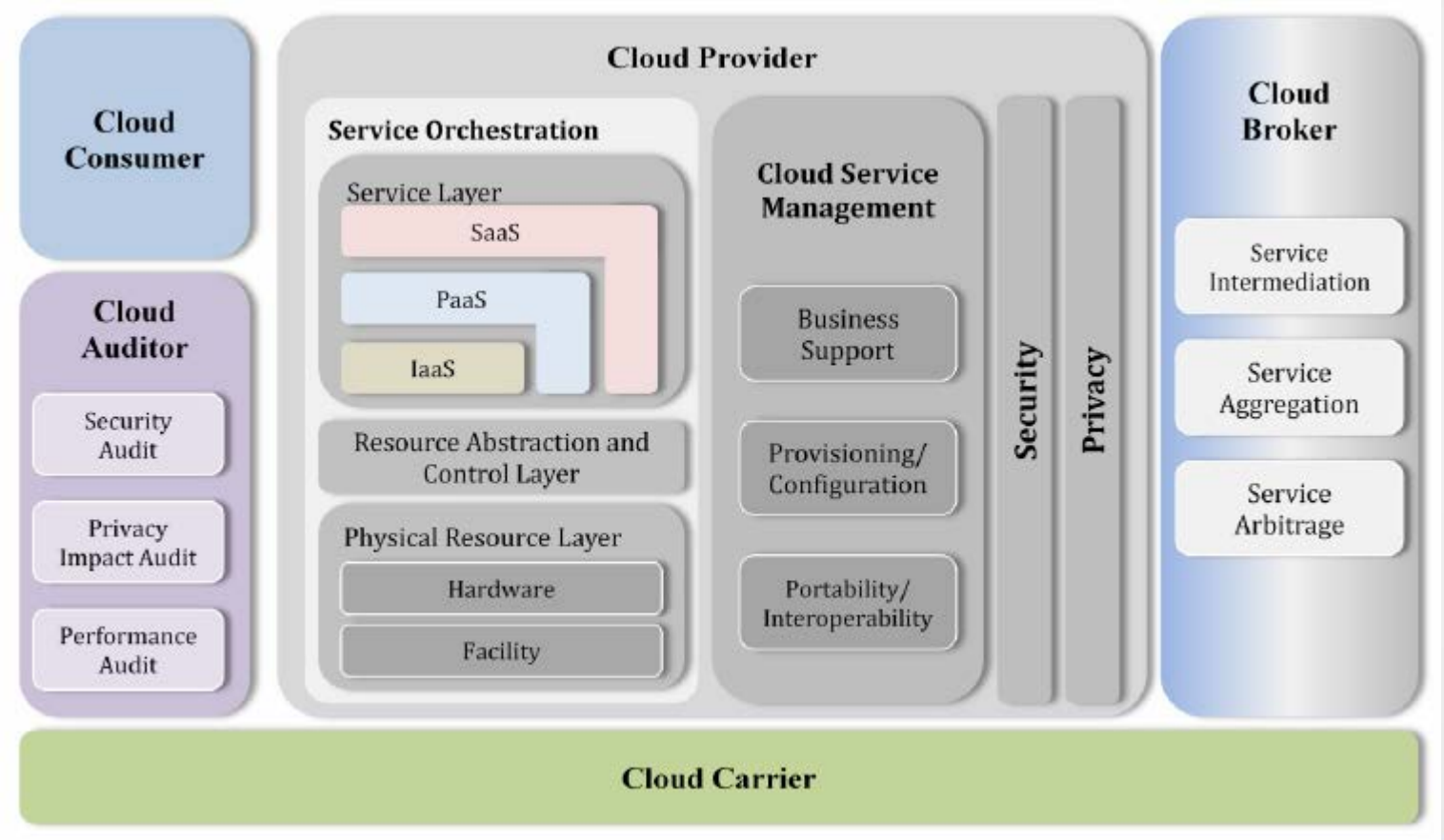



Image source: http://www.softroots.com/public/library/cloud.cfm

# Cloud Computing Services

- IaaS: Infrastructure as a Service
  - Virtual machines, hardware and software
- PaaS: Platform as a Service
  - Google API, web application services
- SaaS: Software as a Service
  - Dropbox, Gmail
- DaaS: Data as a Service
  - TopoLens, data.gov

# Essential Characteristics of Cloud Computing

- On-demand self-service
  - Automatic service for customers
- Broad network access
  - Across different systems (e.g., desktops, mobile phones)
- Resource pooling
  - Consolidating different types of computing resources
- Rapid elasticity
  - Quickly provisioning, allocating, releasing computing resources
  - Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand.
- Measured service
  - Supports “pay-as-you-go” service. Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.
  - Automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service.

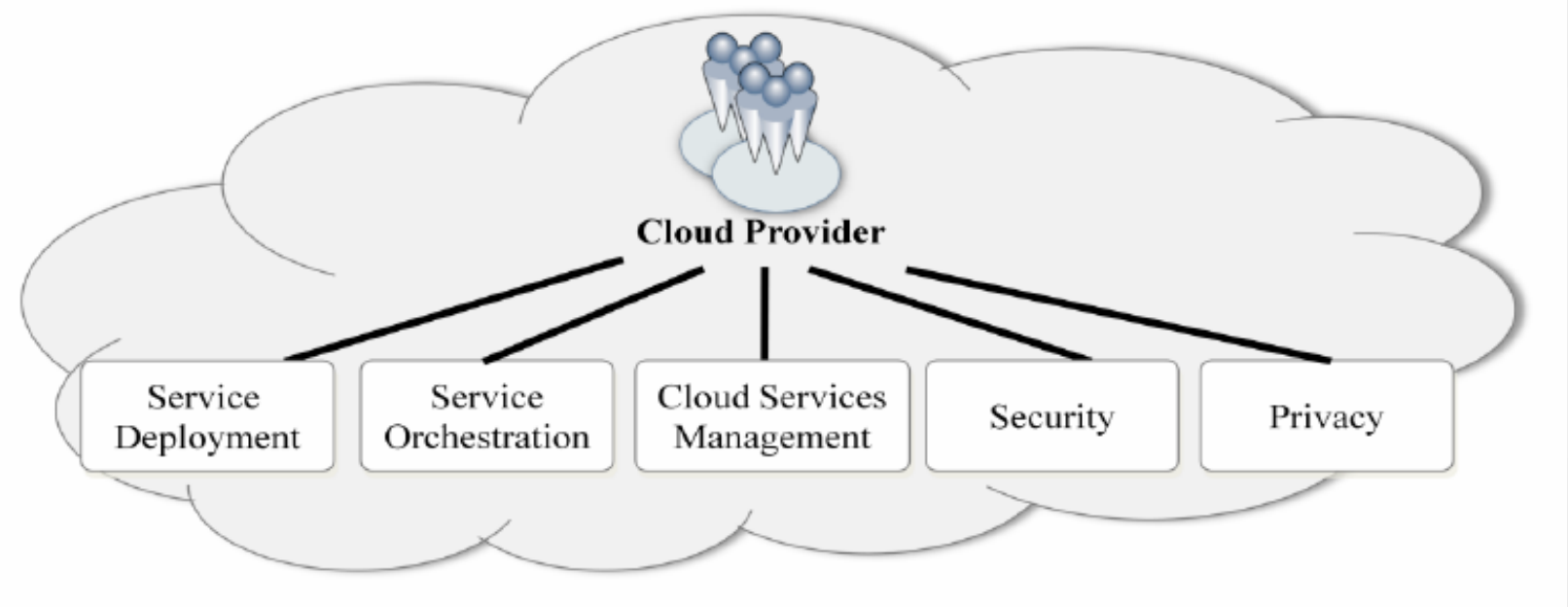
# Cloud Consumer

- SaaS
  - Directly use software or provide members with access
  - Document management, HR, billing, social networks
- PaaS
  - Develop, test, deploy, manage applications
  - Databases, business intelligence, applications
- IaaS
  - Access to virtual computers, storage, network infrastructure
  - Used for backup/recovery, storage, platform hosting

# Cloud Provider

•Acquires computing infrastructure, runs cloud software, delivers services through network.

- SaaS
  - Deploys, configures, maintains software applications
  - Responsible for most of application control, management
- PaaS
  - Manages infrastructure, runs cloud software
  - Supports application deployment, development through software development kits (SDKs)
- IaaS
  - Acquires physical computing resources, runs cloud software (host OS, storage devices, network equipment, etc.)
  - Consumer takes over more fundamental computing resources like OS and network


Cloud Provider
Service Deployment
Service Orchestration
Cloud Services Management
Security
Privacy

# Cloud Auditor

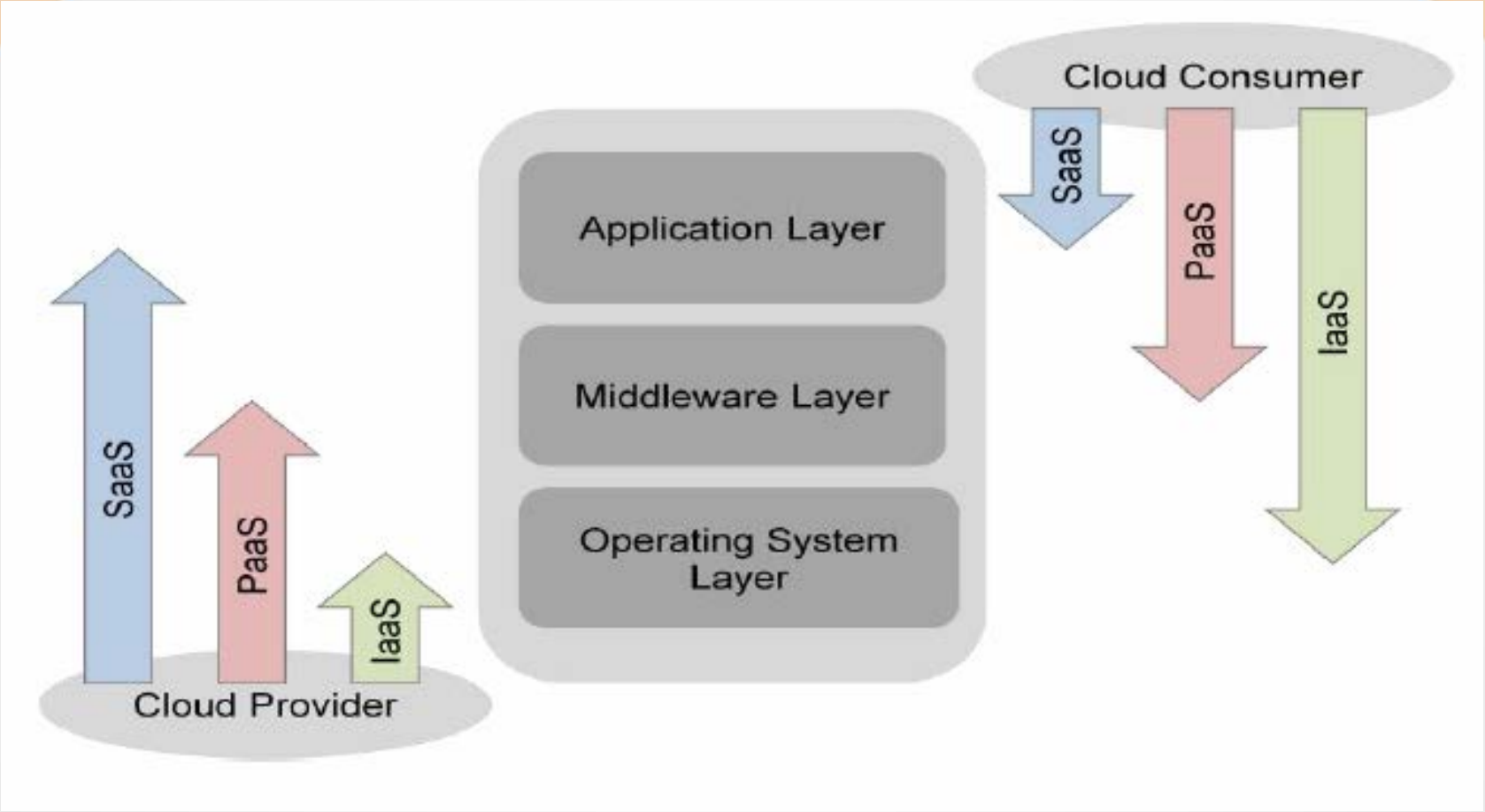
- Evaluate in terms of security controls, privacy impact, performance, etc.
- Ensure confidentiality, integrity, and availability

# Cloud Broker

- An entity that manages the use, performance and delivery of cloud services
- Able to negotiate relationships between cloud providers and cloud consumers
- Provide three category of services
  - Service Intermediation
  - Service Aggregation
  - Service Arbitrage

# Cloud Carrier

- Provides connectivity and transport of cloud services between cloud consumers and cloud providers.


Cloud Consumer
SaaS
PaaS
IaaS
Application Layer
Middleware Layer
Operating System Layer
SaaS
PaaS
IaaS
Cloud Provider

# Geospatial Cloud Services

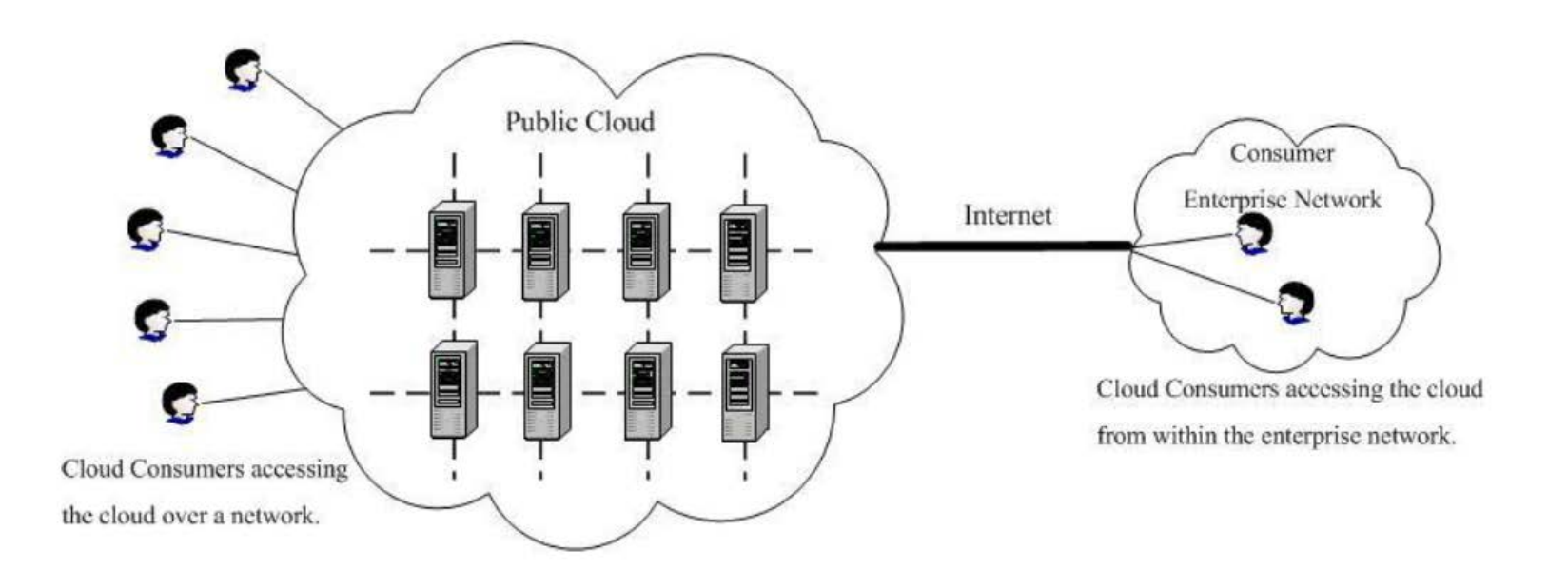
- IaaS
  - Geospatial modeling
  - Allows for control over computing resources
- PaaS
  - Best for parameter extraction
  - Examples: Vegetation index, sea surface temperature
- SaaS
  - ArcGIS Online
  - Knowledge and decision support
  - Used by experts, managers, or public
- DaaS
  - Earth Observation data access, storage and processing

# Types of clouds

- Public Cloud
- Private Cloud
- Community Cloud
- Hybrid Cloud

# Public Cloud

- The cloud infrastructure and computing resources are made available to the general public over the Internet.

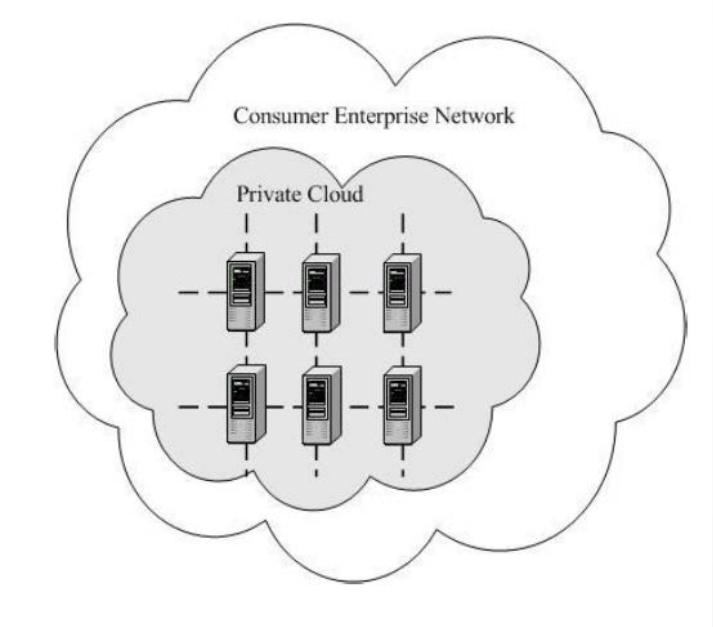


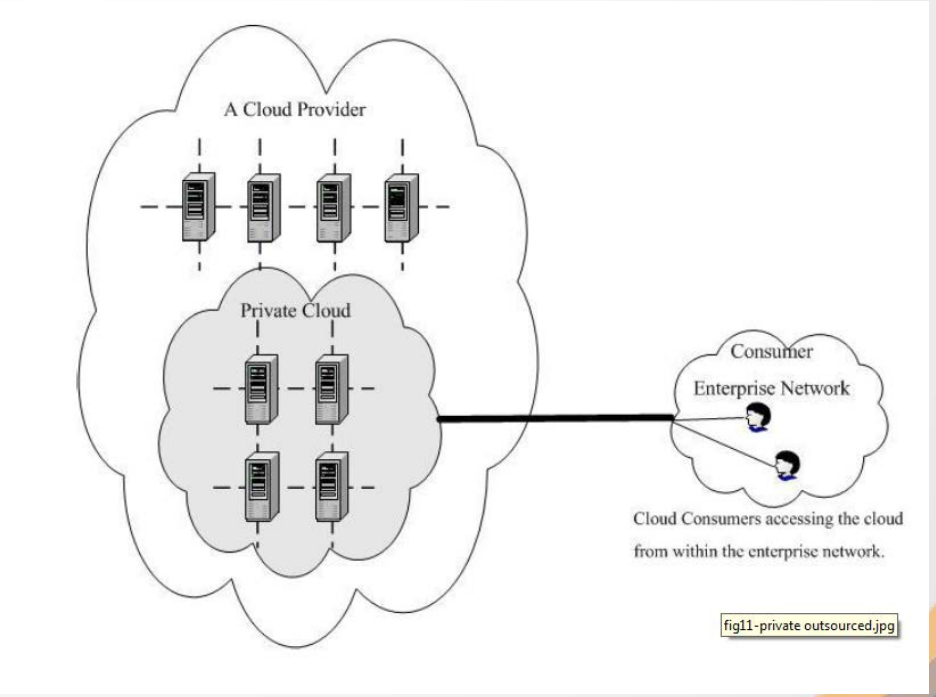
# Public Cloud

- Amazon Elastic Compute Cloud (EC2)
  - http://aws.amazon.com
- IBM's Blue Cloud
- Sun Cloud
- Google AppEngine
- Windows Azure Services

# Private cloud

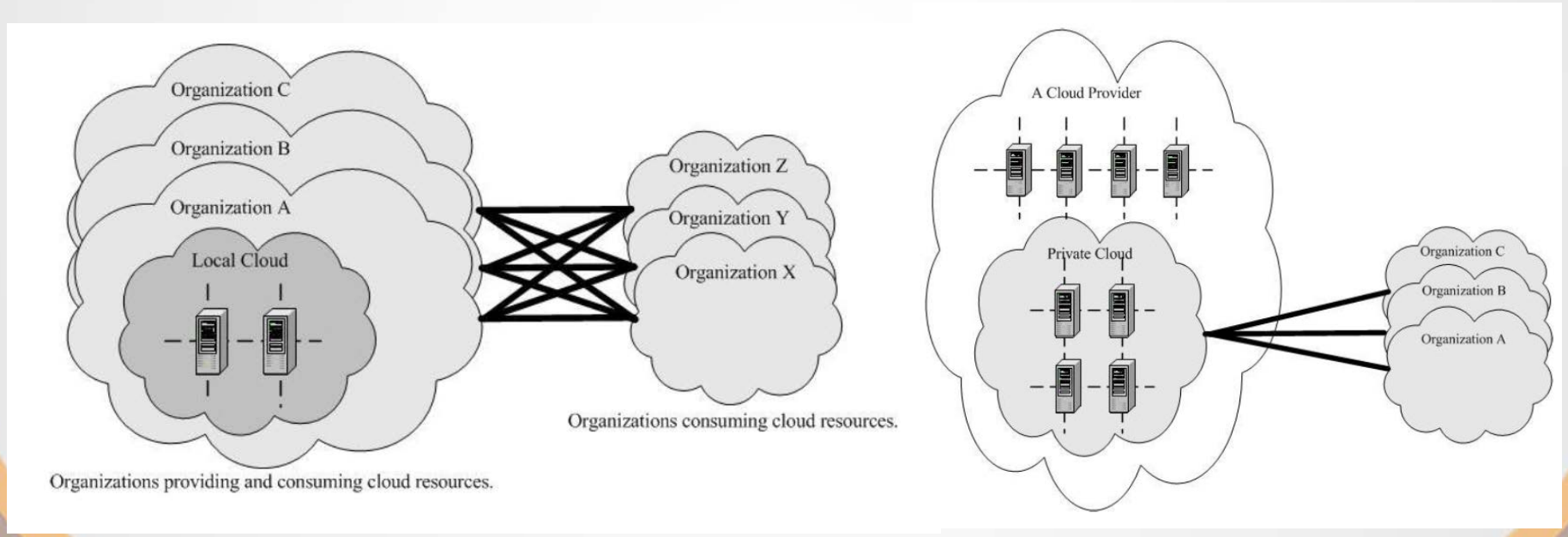
- For a single Cloud Consumer's organization
- Exclusive access given to usage of the infrastructure and computational resources.



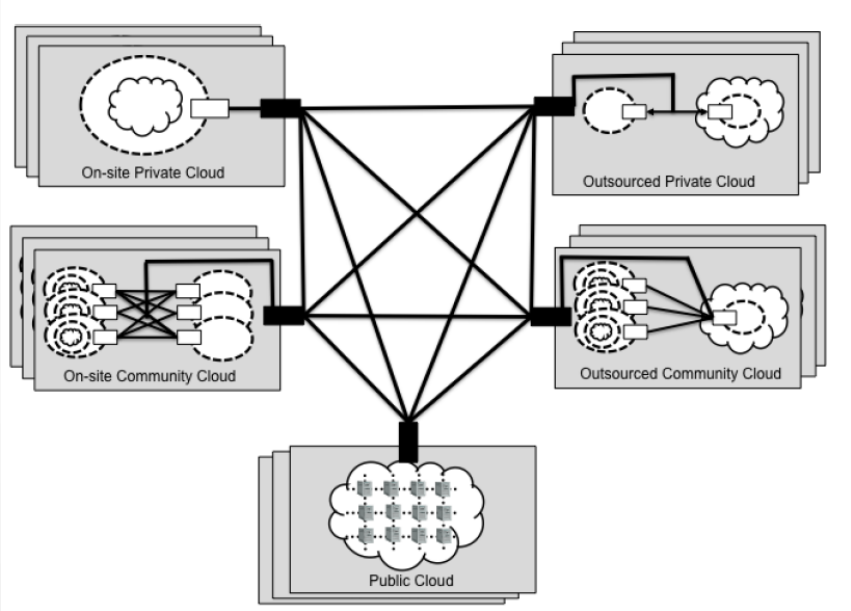
# Community Cloud

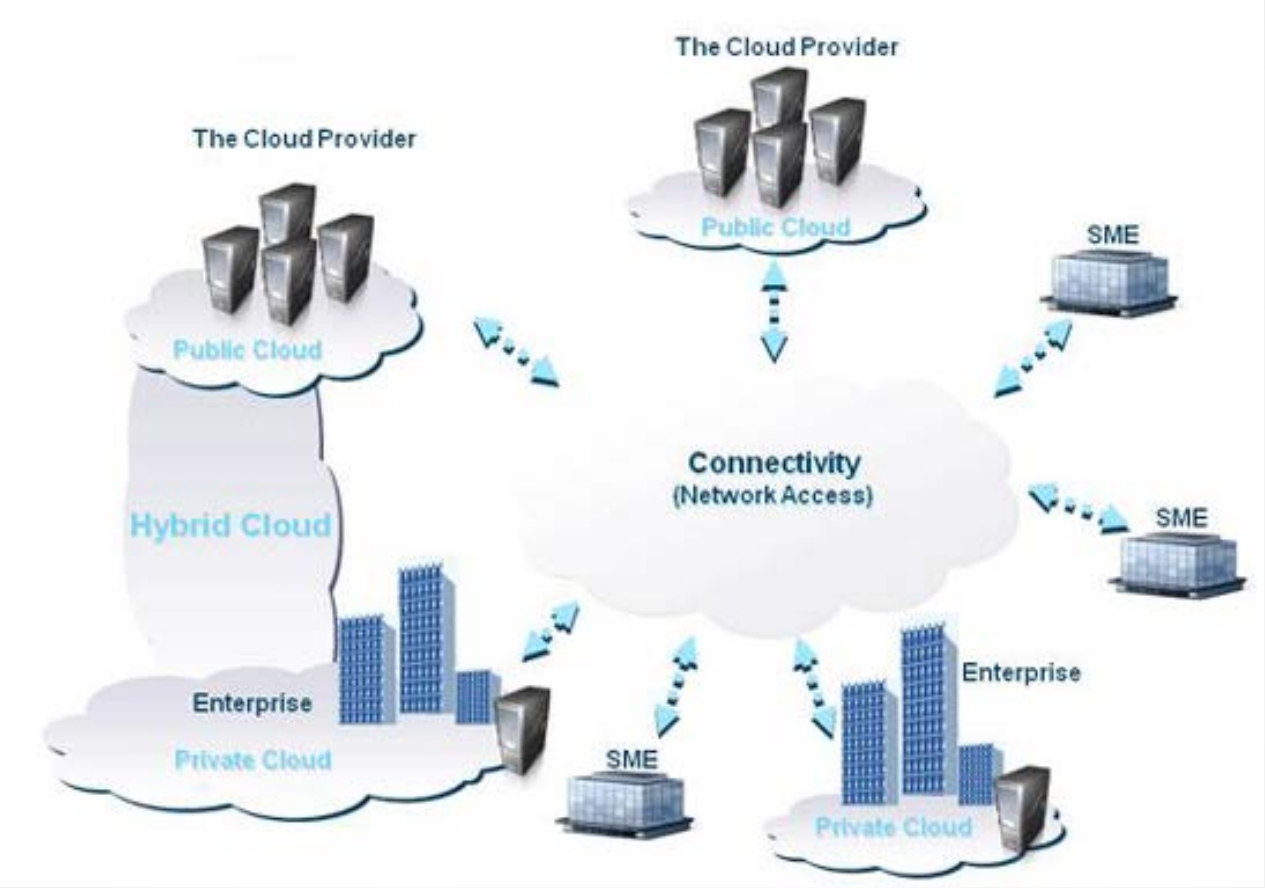
- A group of Cloud Consumers which have shared concerns (e.g. mission objectives, security, privacy and compliance policy)

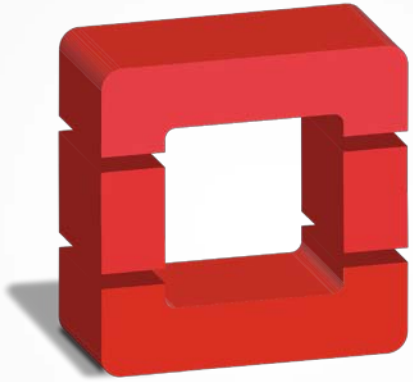


# Hybrid Cloud

- A composition of two or more clouds
- These clouds are bound together by standardized or proprietary technology that enables data and application portability.




The Cloud Provider
The Cloud Provider
Public Cloud
Public Cloud
SME
Connectivity
(Network Access)
Hybrid Cloud
SME
Enterprise
Private Cloud
SME
Enterprise
Private Cloud



openstack®

CLOUD SOFTWARE

# What is OpenStack?

A cloud operating system that controls large pools of computing, storage, and networking resources.

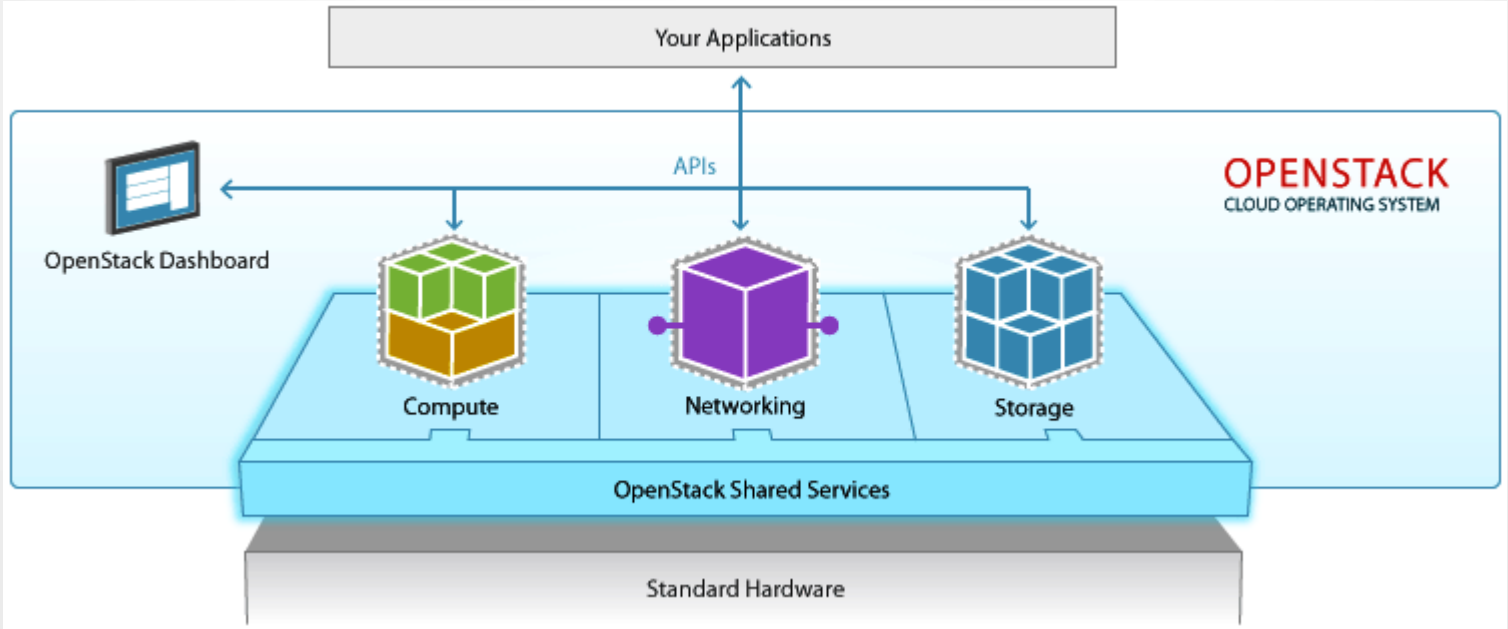



image source: openstack.org

# Benefits of OpenStack

**Flexibility** – immediately create new virtual servers with user specified computing, storage, networking, and OS requirements

**Manageability** – monitor existing servers, reboot or suspend existing servers, change storage, create complete system images as backups

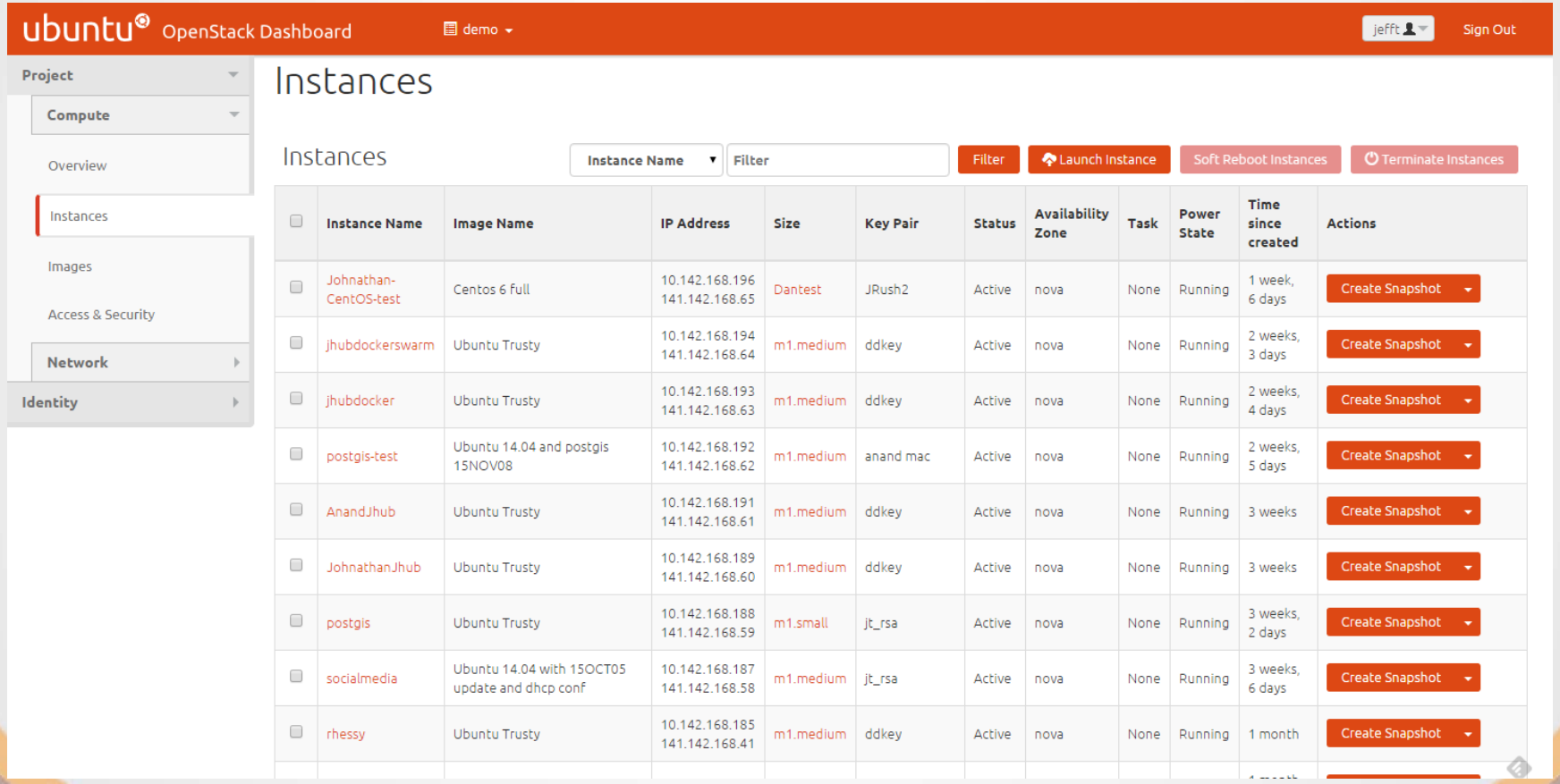
**Sharing and Collaboration** – create a duplicate server for development or to isolate groups, share a system image with pre-installed software collaborators

Instance – a virtual server running in a cloud environment
Snapshot – a backup of an instance or storage block
Image – a bootable snapshot that contains an operating system

# OpenStack Components

- **Nova** – cloud computing controller which manages computing resources and instances
- **Glance** – operating system image service
- **Swift** – object storage for archiving and sharing large amounts of data
- **Keystone** – identity service and access control of OpenStack services
- **Neutron** – networking and IP address management
- **Cinder** – block storage management for managing file system storage
- **Heat** – provides orchestration where multiple composite cloud implementations

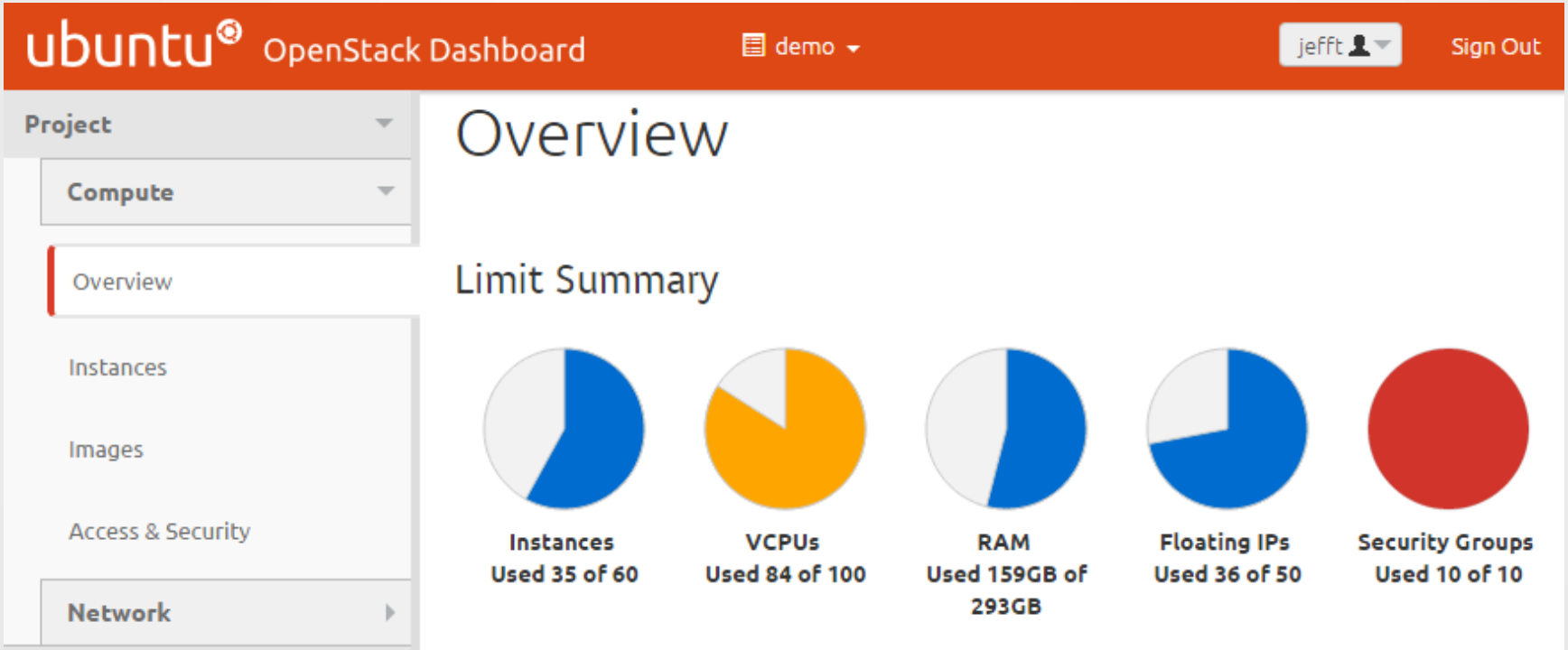
# Horizon the OpenStack Dashboard



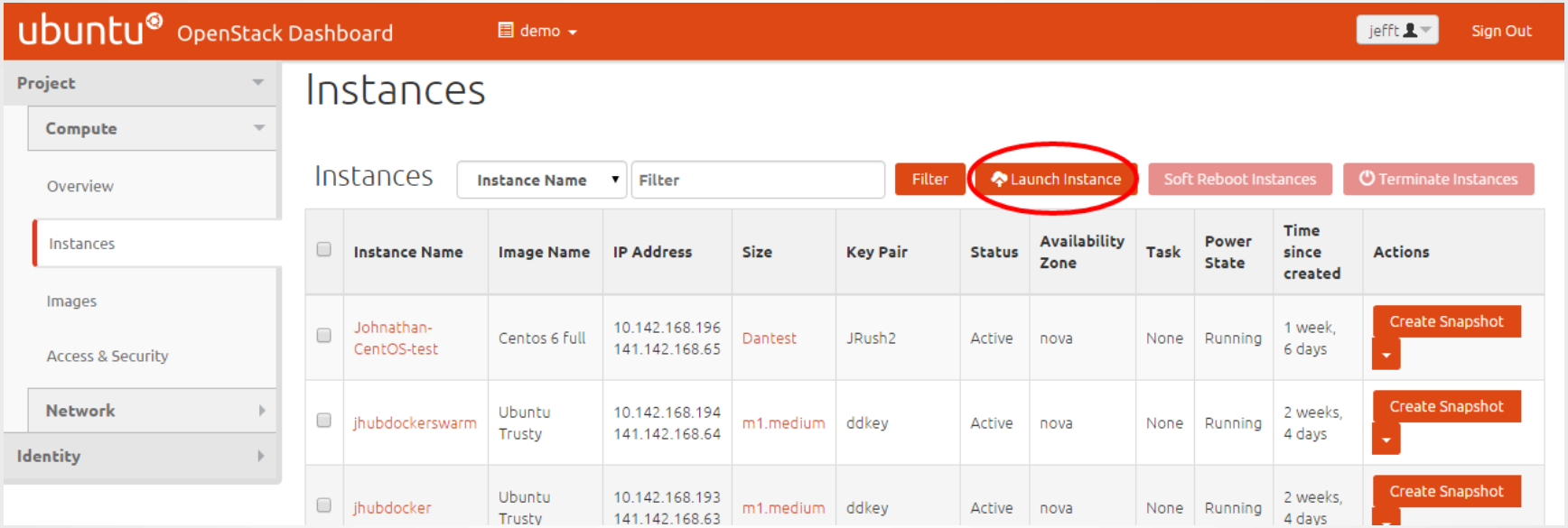
# OpenStack on ROGER

- Currently using 4 nodes each with 20-cores and 128GB of memory
- Shared access to the 4.5PB of storage on ROGER
- 35 active instances
  - o web servers
  - o geoserver and database services
  - o gateway apps
  - o development

# Current Utilization



# Create New Instance



# Available Machine Sizes

| Flavor | CPUs | Memory | Disk |
|---|---|---|---|
| r1.micro | 1 | 512 MB | 5 GB |
| m1.small | 1 | 2 GB | 20 GB |
| m1.medium | 2 | 4 GB | 40 GB |
| m1.large | 4 | 8 GB | 80 GB |
| JR.large | 10 | 16 GB | 100 GB |

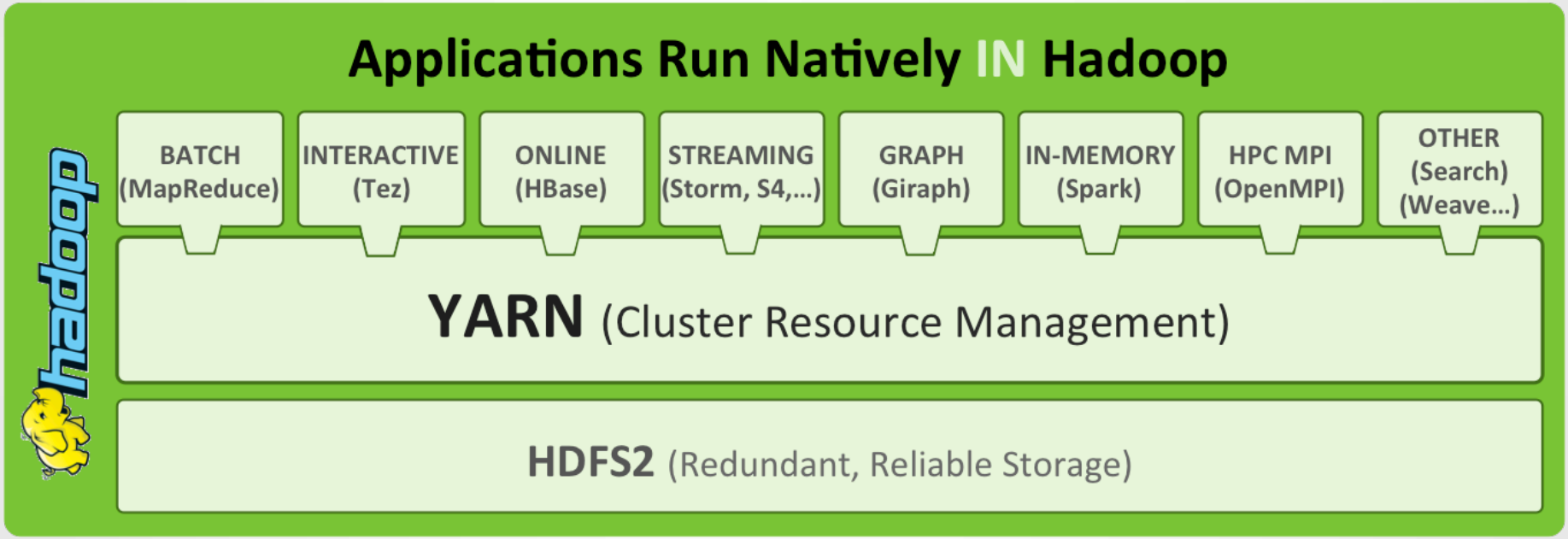
# Available Images

- Base Operating Systems
  - Centos 6
  - Centos 7
  - Fedora 20
  - Ubuntu Trusty
  - Microsoft Windows WS2012R2
- Custom Environments
  - PostgreSQL/PostGIS 9.3
  - Jupyter Hub
  - GitHub Enterprise 2.2.5
- Server Snapshots

# Distributed Computing using Hadoop

- What is Hadoop
  - Distributed file system + replicating data in multiple nodes.
  - Easy-to-use MapReduce interface.
  - Scalable.
  - Parallel execution of mappers/reducers.
  - Fault tolerant.
- When to use Hadoop
  - Data is too big to fit into memory.
  - Tasks can be decomposed as batch-based processing.
  - The problem can be modeled by MapReduce computing paradigm.
  - Scalability is the major issue not interactivity.

# Hadoop Architecture

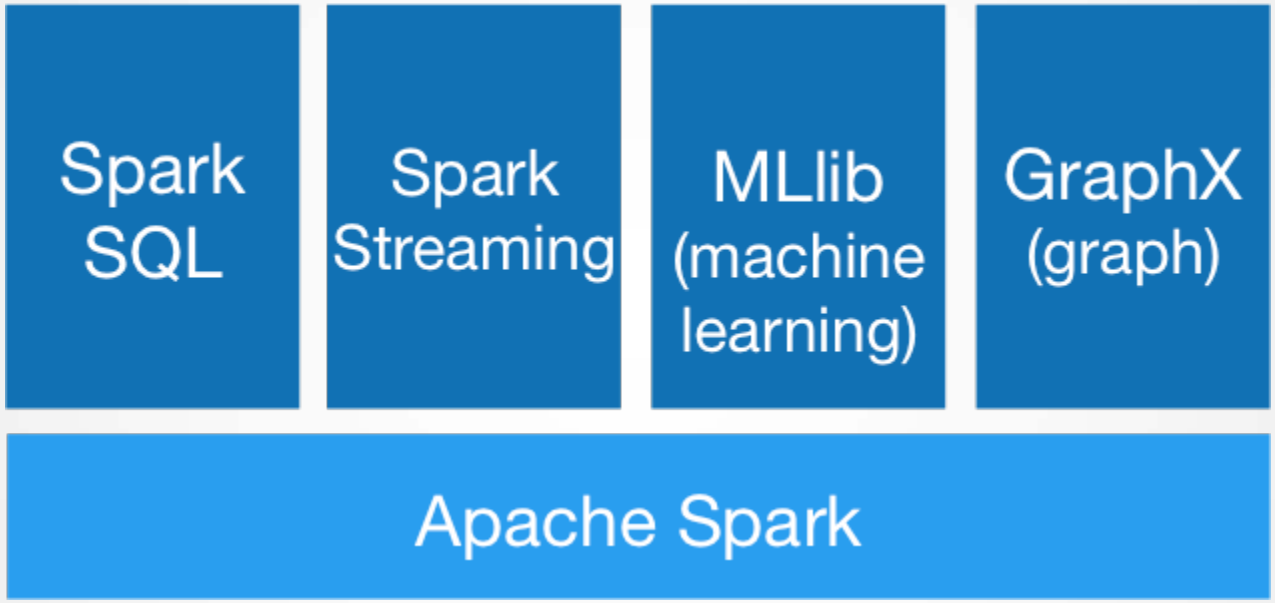



Source: http://radar.oreilly.com/2014/01/an-introduction-to-hadoop-2-0-understanding-the-new-data-operating-system.html

# Distributed Computing using Spark

- What is Spark
  - An open-source cluster computing framework originally developed in the AMPLab at UC Berkeley
  - The fundamental programming abstraction is Resilient Distributed Datasets (RDD), which is a logical collection of data partitioned across machines.
  - Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.
  - Ease of Use
    - Write applications quickly in Java, Scala, Python, R.
  - Runs Everywhere
    - Spark runs on Hadoop, Mesos, standalone, or in the cloud. It can access diverse data sources including HDFS, Cassandra, HBase, and S3.
- When to use Spark
  - When the operations involves iterations, especially machine learning and data mining algorithms
  - Repeatable/multiple queries on the same large dataset

# Spark




Source: http://spark.apache.org/

# Useful resources and links

- XSEDE training
  - https://portal.xsede.org/training/overview
  - https://www.xsede.org/web/xup/online-training
- Free Azure training March 18th, 2016
  - http://research.microsoft.com/en-US/events/azure4researchtraining-illinois2016/default.aspx
  - **Date**: March 18, 2016
    **Time**: 9:00 A.M. to 5:00 P.M. (check-in begins at 8:30 A.M.)
    **Location**: University of Illinois
    **Building/room**: 2405 Siebel Center
    **Address**: 201 N Goodwin Ave, Urbana, Illinois 61807-2302
  - Lunch will be provided

# cyberinfrastructure

CyberGIS is Geographic Information Science and Systems based on advanced cyberinfrastructure.

Cyberinfrastructure includes:

- (high performance) computing systems
- data storage systems
- advanced instruments
- data repositories
- visualization environments
- people
- linked by high speed networks

# CyberInfrastructure Resources

NSF Blue Waters @ UofI
13,300 TFlop/s

NSF XSEDE: SDSC Comet
2,000 TFlop/s

UofI Campus Cluster

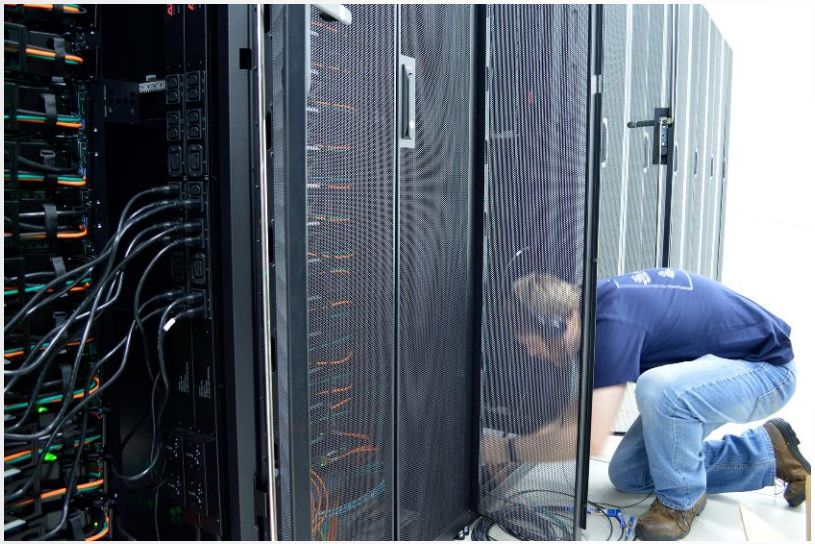
# CyberGIS ROGER

**CyberGIS ROGER**
Resourcing Open Geospatial Education and Research
After Roger Tomlinson
goo.gl/8MpYC9
~60 TFlop/s

# Why ROGER?

- Both a research project and a research resource
- Configured to best support geospatial data, with emphases on local memory and shared storage size and speed
- Supports multiple paradigms: traditional batch HPC, Hadoop, and Cloud (OpenStack).
- Integrate the three paradigms and leverage their strengths
- Inform the design of future geospatial supercomputers

# ROGER: Overview



- Provides the CyberGIS Center with HPC, Hadoop and OpenStack functionality all on one system
- Managed by Systems Group

  With support from our storage, network, security and services groups

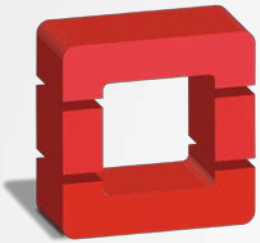
# ROGER: File system

- IBM General Parallel File System (GPFS)
- Parallel file system means all of the nodes get fast access to hard drive storage.

# ROGER for this class

- ROGER accounts for every student in this class
- Interact with various components in ROGER
  - Job submission
  - Handling and processing massive data with Hadoop, Spark, R, and Python, etc.
  - Getting to know OpenStack
  - Working with Geospatial libraries

# Discussion on ROGER Resources

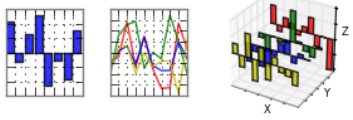


openstack
CLOUD SOFTWARE


mongoDB


pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



# Register to ROGER

1. Provide your netID to the Google Doc:

https://docs.google.com/spreadsheets/d/1zWkT6RnUiwvcYzEBIKCNHUIZFx0G-dJLBb_mmI7r7Ys/edit?usp=sharing

2. After the lecture, the above document will not be accessible (for privacy concerns)

# Reference


- Liu, F., Tong, J., Mao, J., Bohn, R., Messina, J., Badger, L., & Leaf, D. (2011). NIST cloud computing reference architecture. NIST special publication, 500, 292.
- Yang, C., Raskin, R., Goodchild, M., & Gahegan, M. (2010). Geospatial cyberinfrastructure: past, present and future. *Computers, Environment and Urban Systems*, *34*(4), 264-277.


- Some slides were derived from CyberGIS Fellows:
  - Jie Tian Clark University
  - Wenwen Li, Arizona State University
  - Yi Qiang and Nina Lam, Louisiana State University