



Copyright 2012 California Institute of Technology. Government sponsorship acknowledged.

# Image Processing on the Cloud

Emily Law

Cloud Computing Workshop  
ESIP 2012 Summer Meeting  
July 14<sup>th</sup>, 2012





# Outline

- Cloud computing @ JPL SDS
- Lunar images
- Challenge
- Image tiling process
- Implementations
- Analysis
- Summary



# Science Data Systems

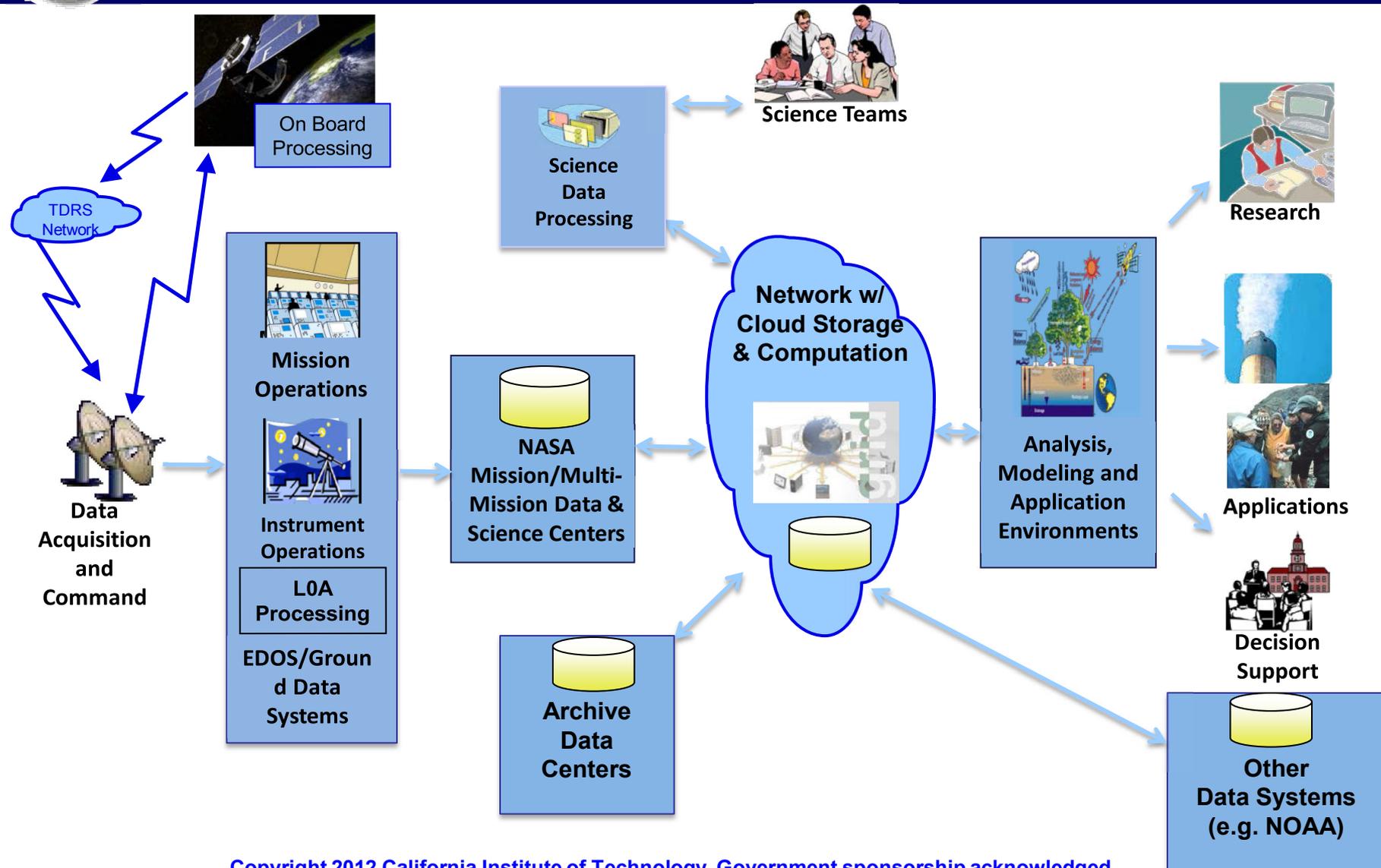
- Cover a wide variety of domain disciplines
  - Solar system exploration, Astrophysics, Earth science, Biomedicine, etc,...
- Each has its own communities, standards and systems
- But, there is a set of common components & constraints
- Some can greatly benefit from proven cloud computing technology





# Earth Science Data Systems

CALIFORNIA INSTITUTE OF TECHNOLOGY



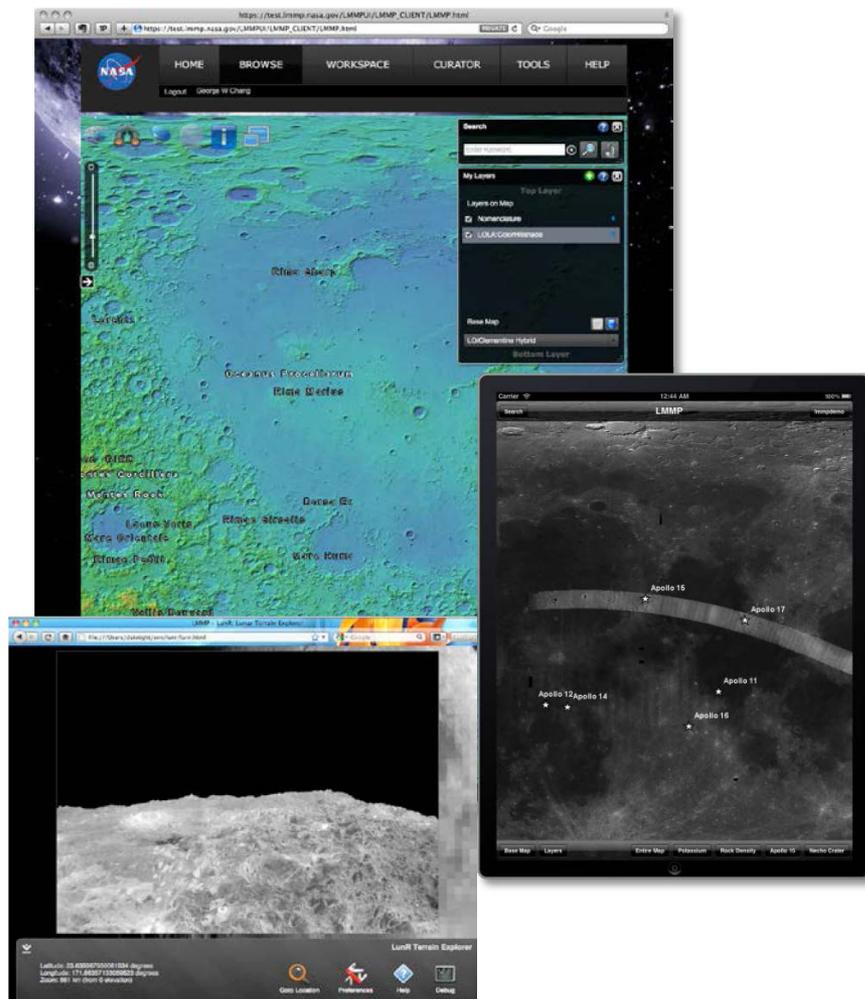
Copyright 2012 California Institute of Technology. Government sponsorship acknowledged.



# Lunar Modeling and Mapping Project (LMMP)

CALIFORNIA INSTITUTE OF TECHNOLOGY

- Provides science and exploration community a suite of lunar mapping and modeling tools and products that support the lunar exploration activities
- The tools and products are made available through a common, intuitive NASA portal
- Utilizes open standards and facilitates platform and application independent access





# Challenge

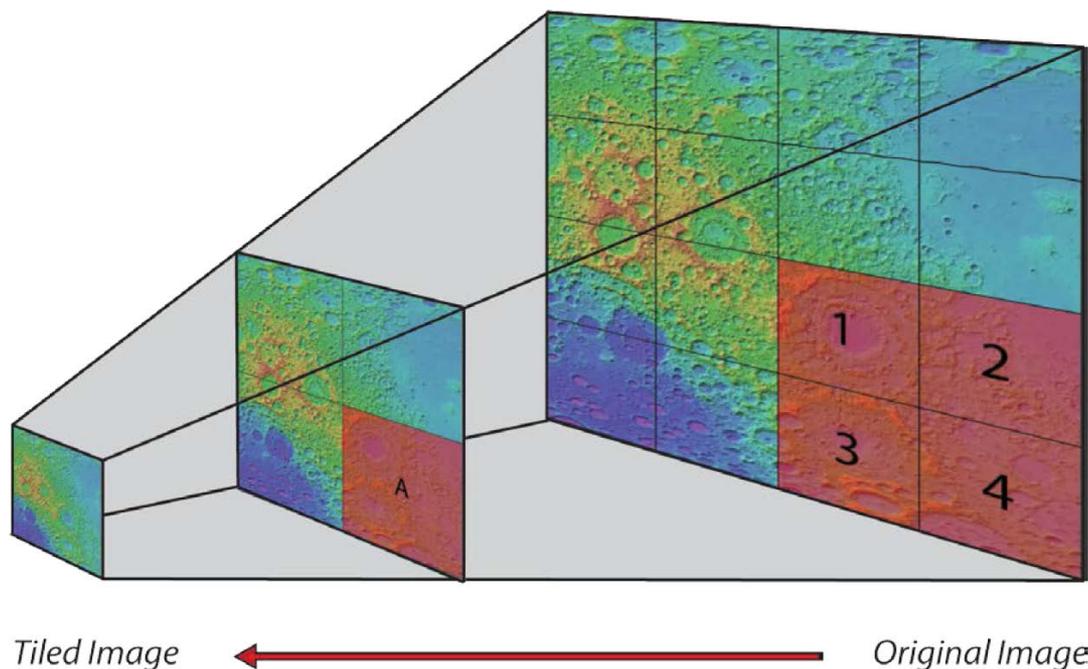
- How to make these large images usable by desktop computers, mobile devices and other memory constrained products?





# Tiling Process

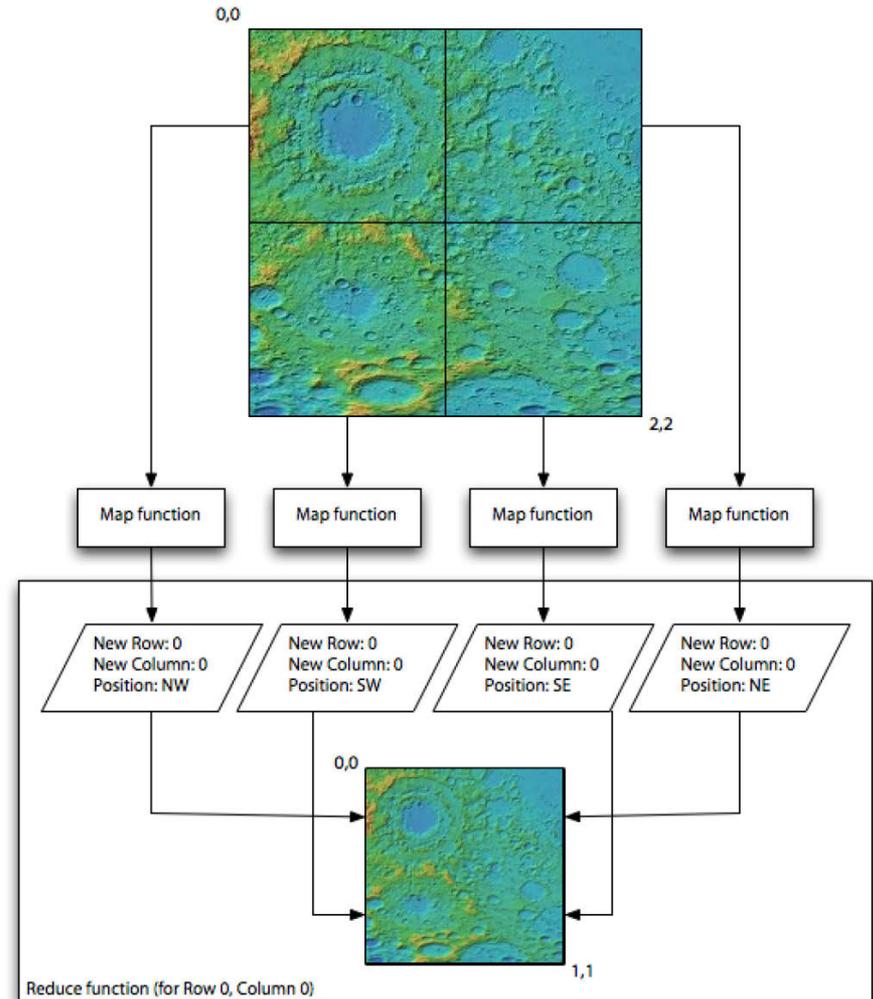
- Divides images into small tiles
- Combines and shrinks for the next zoom level
- Iterates till the zoom level has only 1 tile





# Using Hadoop

- Hadoop is an implementation of Google's Map-Reduce algorithm
- *Map* Function – Takes a subset of the data, performs a computation, and returns an output.
- *Reduce* Function – Consolidates outputs from the *map* function to generate another output





# In-House Implementation

- Test image, 2.77 gigabytes LRO LOLA (Lunar Orbiter Laser Altimeter) colorized digital elevation map which produced 9.1 gigabytes set of tiles
- Ran Hadoop on local machines in the lab
- 2 Sun Fire x4170 machines running dual Xeon X5570 processors with 72 GBs of RAM with a heterogeneous mix of Solaris 10 and Linux
- Performance was excellent
- Machines are costly to maintain, especially since these tasks are “bursty”



# Cloud Implementation Using Amazon EC2

CALIFORNIA INSTITUTE OF TECHNOLOGY

- Amazon EC2 is a cloud computing infrastructure allowing users to “rent” virtual machines
- Installed Hadoop Elastic MapReduce framework on a number of EC2 instances
- Output image files stored on Amazon S3, a cloud storage system

Name	Instance	AMI ID	Root Device	Type	Status	
<input type="checkbox"/>	i-1f5add75	ami-7ea24a17	ebs	cc1.4xlarge	s	
<input type="checkbox"/>	i-3323a459	ami-725fb41b	ebs	cc1.4xlarge	s	
<input type="checkbox"/>	i-3123a45b	ami-725fb41b	ebs	cc1.4xlarge	s	
<input type="checkbox"/>	i-7bb08411	ami-84db39ed	ebs	m1.small	s	
<input type="checkbox"/>	empty	i-01e2df6b	ami-7ea24a17	ebs	cc1.4xlarge	s
<input type="checkbox"/>	lmp-basic-lighting	i-abfc72c7	ami-7a29d913	ebs	m1.large	ru

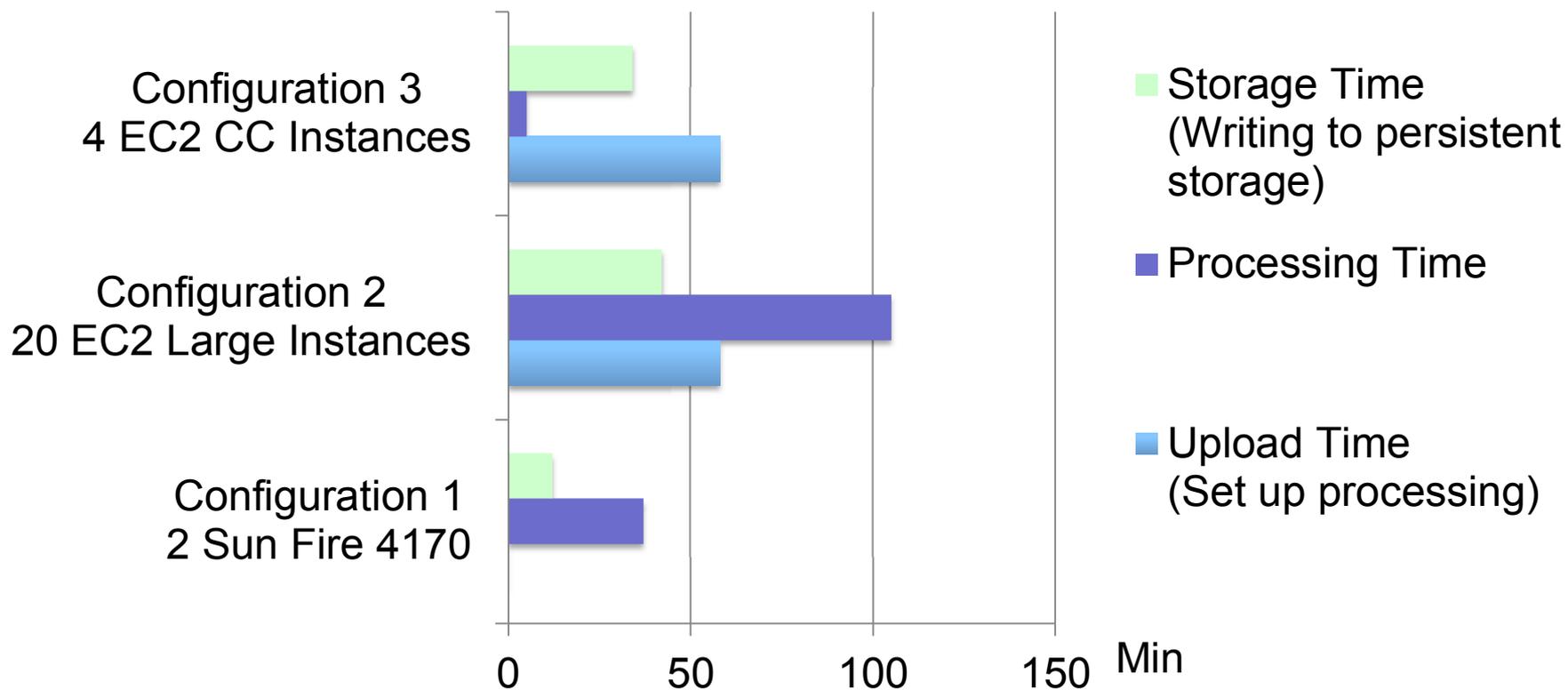


# Configurations

- Configuration 1 - In-House  
2x Sun Fire 4170  
72 GB RAM, 64 GB SSD Storage  
\$10K each, plus administration and infrastructure costs
- Configuration 2 - 20 EC2 “Large”  
20 EC2 Large Instances (4 Compute Units ~ 4x1GHz Xeon)  
7.5 GB RAM, 850 GB Storage  
\$0.34/instance/hour plus bandwidth
- Configuration 3 - 4 EC2 “CC”  
4 EC2 Cluster Compute Instances (33.5 Compute Units)  
23 GB RAM, 1.69 TB Storage  
\$1.60/instance/hour plus bandwidth



# Performance





# Cost

- In-House Implementation
  - **Total Cost: \$20K + SA + infrastructure**
- 20 EC2 Large
  - Processing:  $2\text{h} \times 20 \times \$0.34 = \$13.60$
  - Bandwidth:  $3\text{GB} \times \$0.10 = \$0.30$
  - Storage:  $10\text{GB} \times \$0.14 = \$1.40/\text{month}$
  - **Total Cost: \$15.30**
- 4 EC2 CC
  - Processing:  $1\text{h} \times 4 \times \$1.60 = \$6.40$
  - Bandwidth:  $3\text{GB} \times \$0.10 = \$0.30$
  - Storage:  $10\text{GB} \times \$0.14 = \$1.40/\text{month}$
  - **Total Cost: \$8.10**



# Performance Analysis

## **In-House Implementation**

- Fastest overall
- Did not need to export data to remote systems
- Most expensive from a cost-benefit perspective

## **Cloud Implementation**

- Upload and storage time a consideration
- Network speed between Hadoop nodes a significant consideration
- Most cost-effective for occasional, computationally intensive jobs



# Conclusion

- Hadoop framework provides a simple programmatic interface for developing distributed computing applications for problems that are parallelizable
  - Problems that required large amounts of data will depend on the interconnect speeds between nodes
- Cloud computing gives a cost-effective infrastructure to use compute capacity as needed
- In designing applications for cloud, must consider the performance of locally run machines vs. the price of cloud instances
- Security should also be considered in using public infrastructure
  - We are using a hybrid system where private data is hosted locally while public data is on the cloud