NATIONAL e-SCIENCE
INFRASTRUCTURE CONSORTIUM

# IT Infrastructure for New Paradigm of Science: Contribution to Higgs Particle Discovery

by

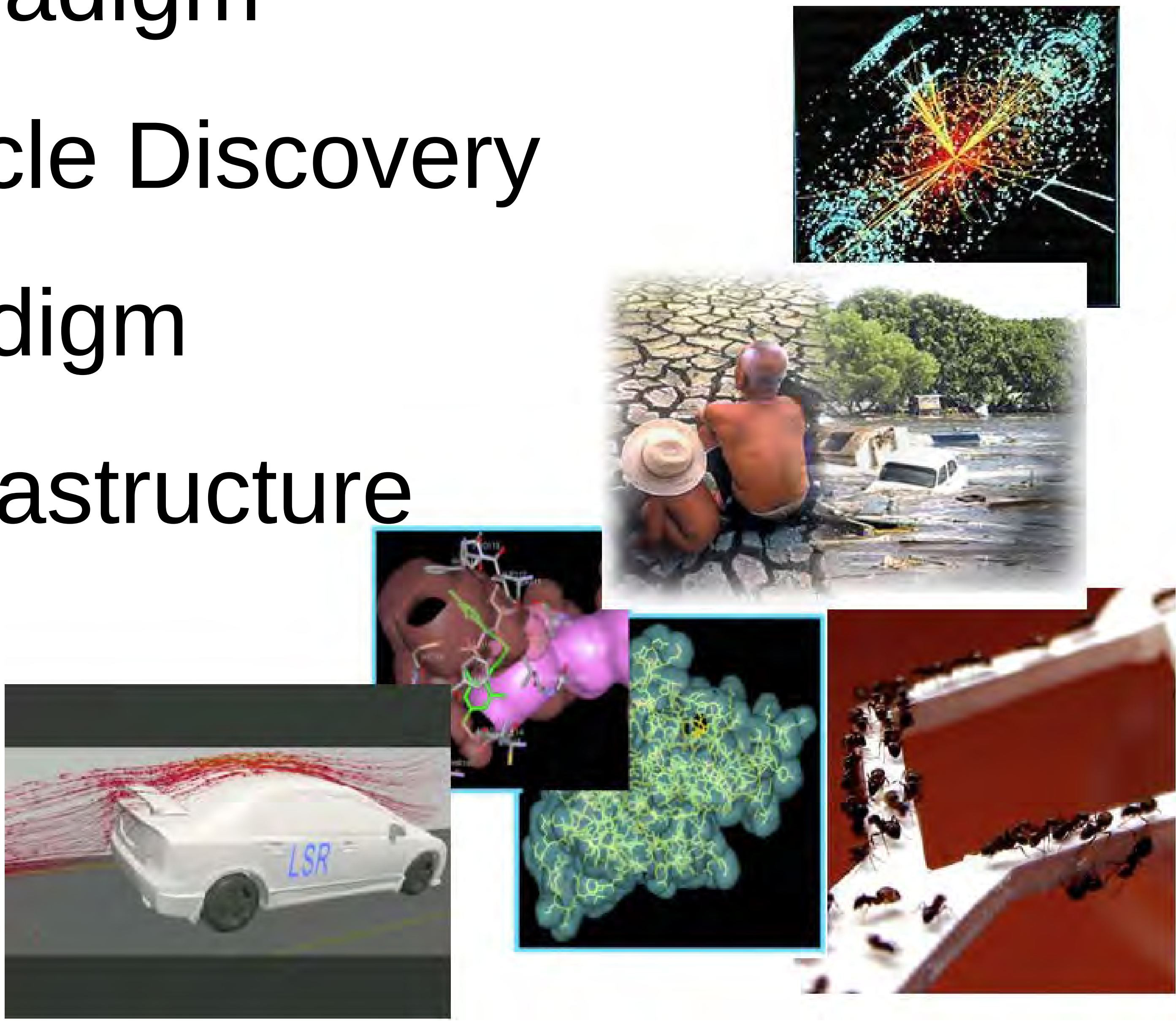
Dr. Burin Asavapibhop
Department of Physics, Faculty of Science,
Chulalongkorn University

Dr. Supakit Prueksaaroon
Large-Scale Simulation Research Laboratory
National Electronics and Computer Technology Center

# Outline

NATIONAL E-SCIENCE INFRASTRUCTURE CONSORTIUM

- Introduction to new Paradigm of Science
- Case Study of new Paradigm
- CERN and Higgs Particle Discovery
- Challenge of new paradigm
- National e-Science Infrastructure Consortium



# Data and Science

- All science is becoming data-driven
  - Physics, Life-science, Bio-technology, …
- Data growing exponentially, in all science
  - Data Collection from: Sensor networks, Satellite surveys, Observation devices, LHC,..
  - Computational intensive analyses results
- Data becoming increasingly open/public/online
  - Regulate to shared/Open the raw data – NSF
  - International distributed collaborations – CERN
  - Open Access Policy for accepted publications – US National Library of Medicine

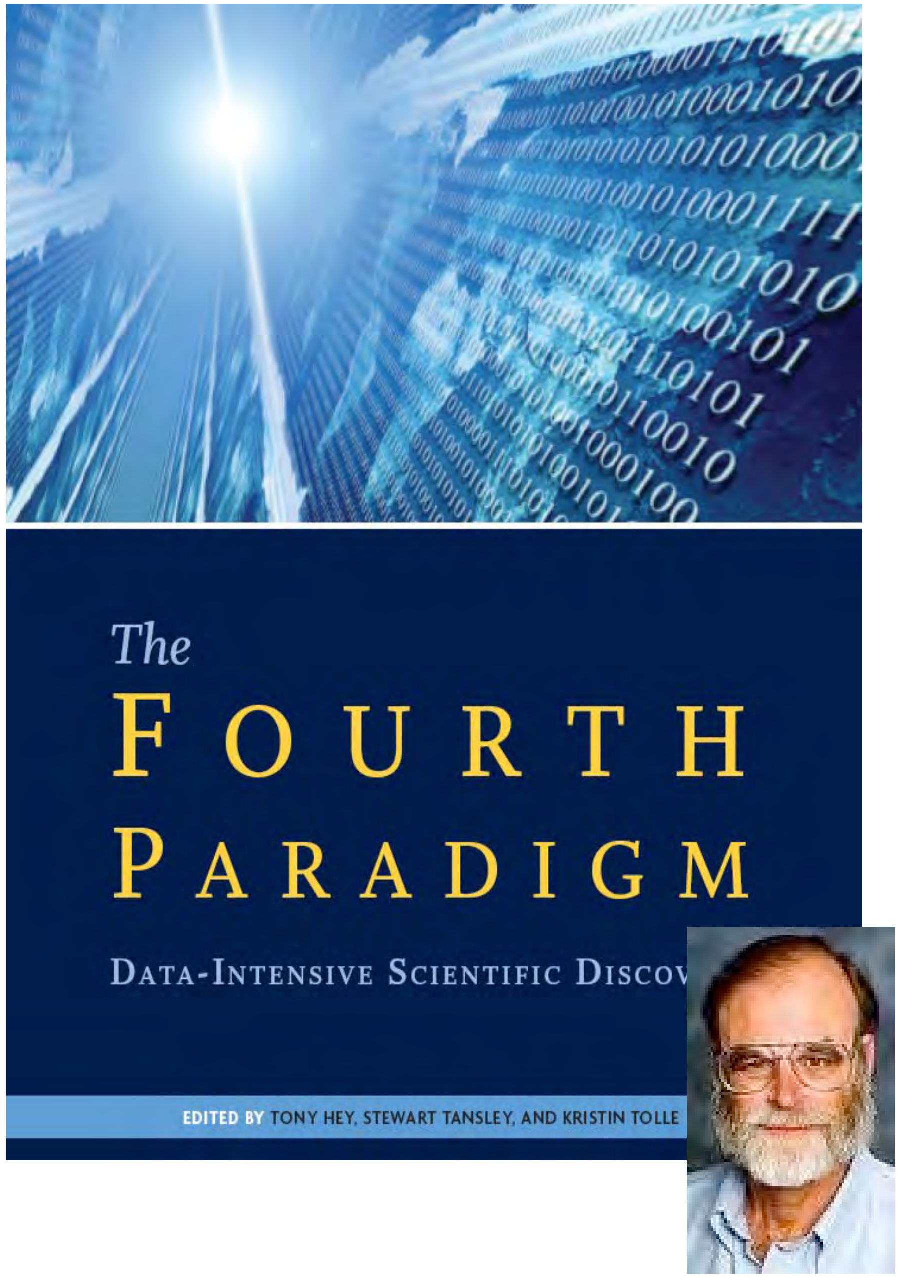
# Data dominate factor

- Data growth rapidly

| | |
|---|---|
| Year 2000 | ~a hundred Gigabytes |
| Year 2005 | ~a few Petabytes |
| Year 2010 | ~a hundred Petabytes |
| Year 2015 | ~a thousand Petabytes? |

Now we're in the new paradigm of scientific research

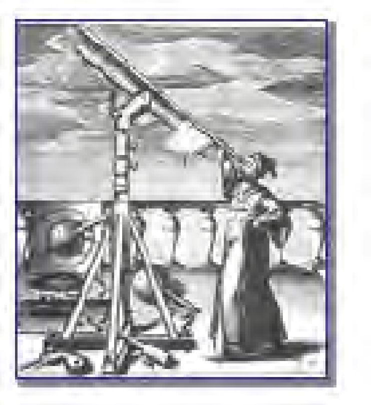
# The Fourth Paradigm


The
Fourth
Paradigm
Data-Intensive Scientific Discov
Edited by Tony Hey, Stewart Tansley, and Kristin Tolle

- Jim Gray
  - Scientific computing is revolving around data
  - Need scale-out solution for analysis
  - Take the analysis to the data!

Propose by: Jim Gray and Tony Hey (Microsoft Research)

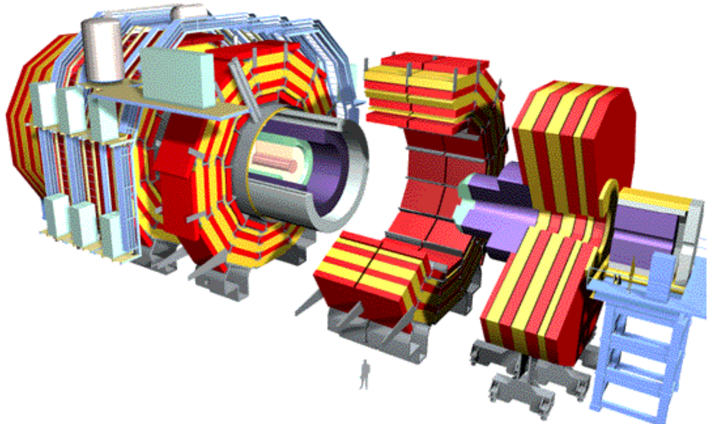
# Emergence of a 4th Research Paradigm



- Thousand years ago – **Experimental Science**
  - Observe natural phenomena and attempt to classify
- Last few hundred years – **Theoretical Science**
  - Describe above classification with mathematical model: Newton's Laws, Maxwell's Equations…

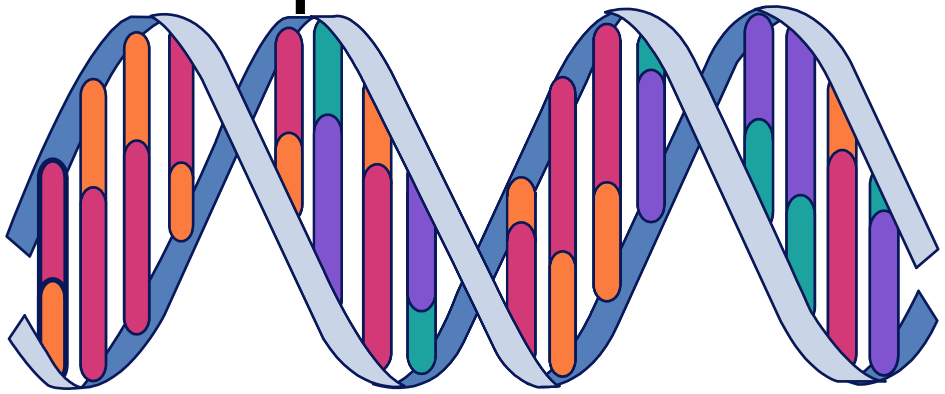
$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

- Last few decades – **Computational Science**
  - Simulation of complex phenomena using above math model
- Today – **Data-Intensive Science**
  - Combine empirical, theoretical, and computational branches with DATA
    - For analysis and data mining
    - For data visualization and exploration
    - For scholarly communication and dissemination



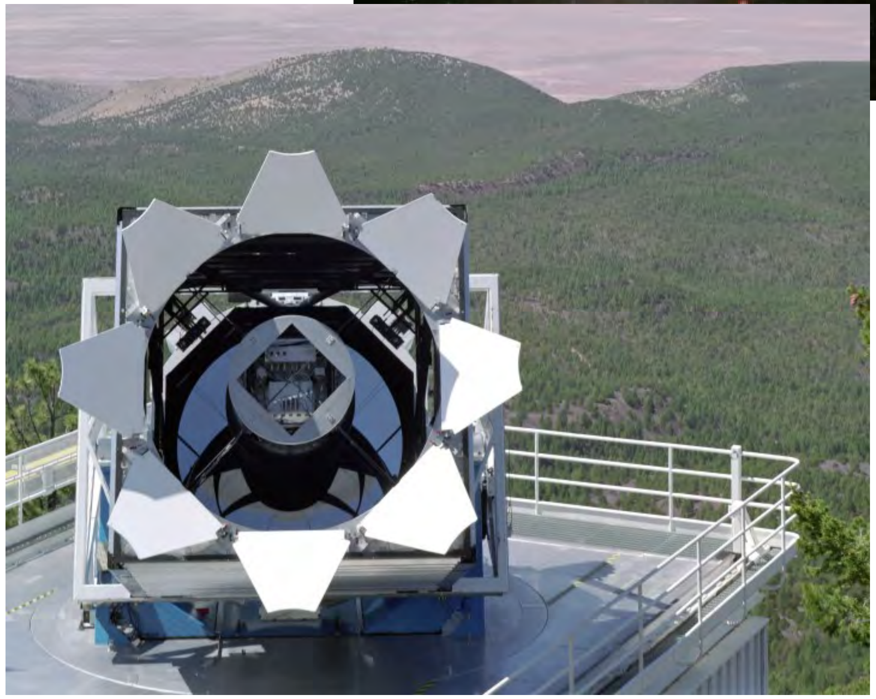
# Data Intensive Science

- Example of data-intensive science
- Astronomy
  - Digital sky surveys
- Molecular Genomics and related disciplines
  - Human Genome, other genome Database
- High energy & nuclear physics
  - Experiment at CERN Large Hadron Collider



# Astronomy

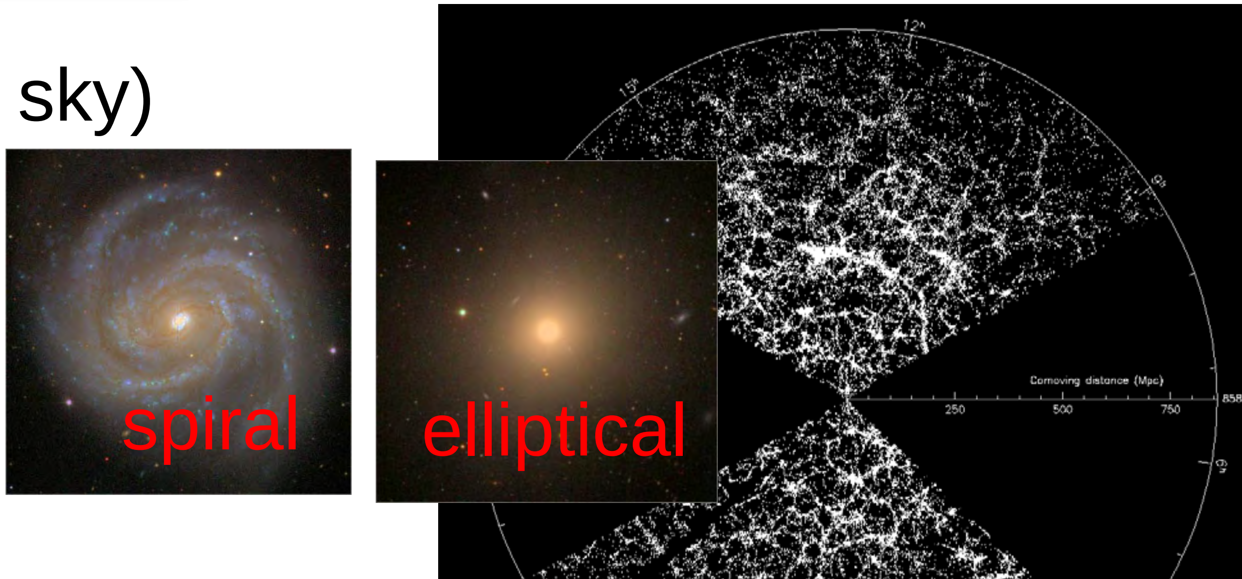
- In the “old days” astronomers took photos.
- New instruments are digital (< 100 GB/night)
  - Detectors are following Moore’s law.
- Data avalanche: double every 2 years
- Why astronomy interesting?
  - Cannot do experiments!!!
  - No privacy restrictions
  - No intellectual property
  - No one wants to sell you data
  - And there’s a lot of it

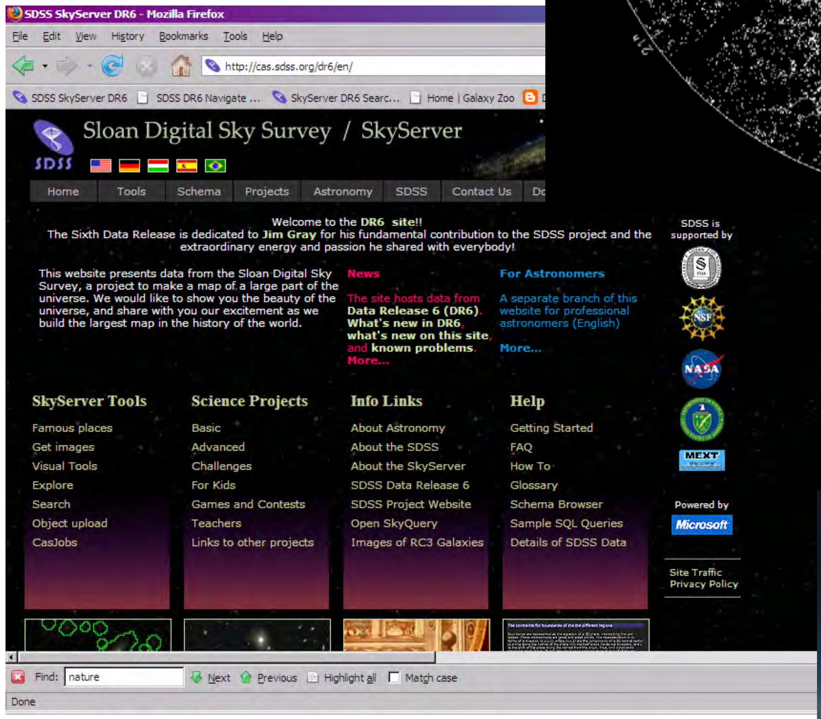


# Sloan Digital Sky Survey

National e-Science Infrastructure Consortium

- Created a map of the Universe (25% of sky)
- Started in 1992, finished in 2008
- 2.5Terapixels of images
- 10TB of raw, 120TB of processed
- Database, **Free to anyone**
- 287 million sky objects
- Results: 1900 Citations
  - 40 million visual galaxy classification by the public
  - 300,000 people participating
  - Original discoveries by the public (Voorwerp, Green Peas)





Welcome to the **DR6** site!!
The Sixth Data Release is dedicated to **Jim Gray** for his fundamental contribution to the SDSS project and the extraordinary energy and passion he shared with everybody!

This website presents data from the Sloan Digital Sky Survey, a project to make a map of a large part of the universe. We would like to show you the beauty of the universe, and share with you our excitement as we build the largest map in the history of the world.

**News**

The site hosts data from **Data Release 6 (DR6)**. **What's new in DR6**, **what's new on this site**, and **known problems**. **More...**

**For Astronomers**

A separate branch of this website for professional astronomers (English)

**More...**

SDSS is supported by

**SkyServer Tools**
Famous places
Get images
Visual Tools
Explore
Search
Object upload
CasJobs

**Science Projects**
Basic
Advanced
Challenges
For Kids
Games and Contests
Teachers
Links to other projects

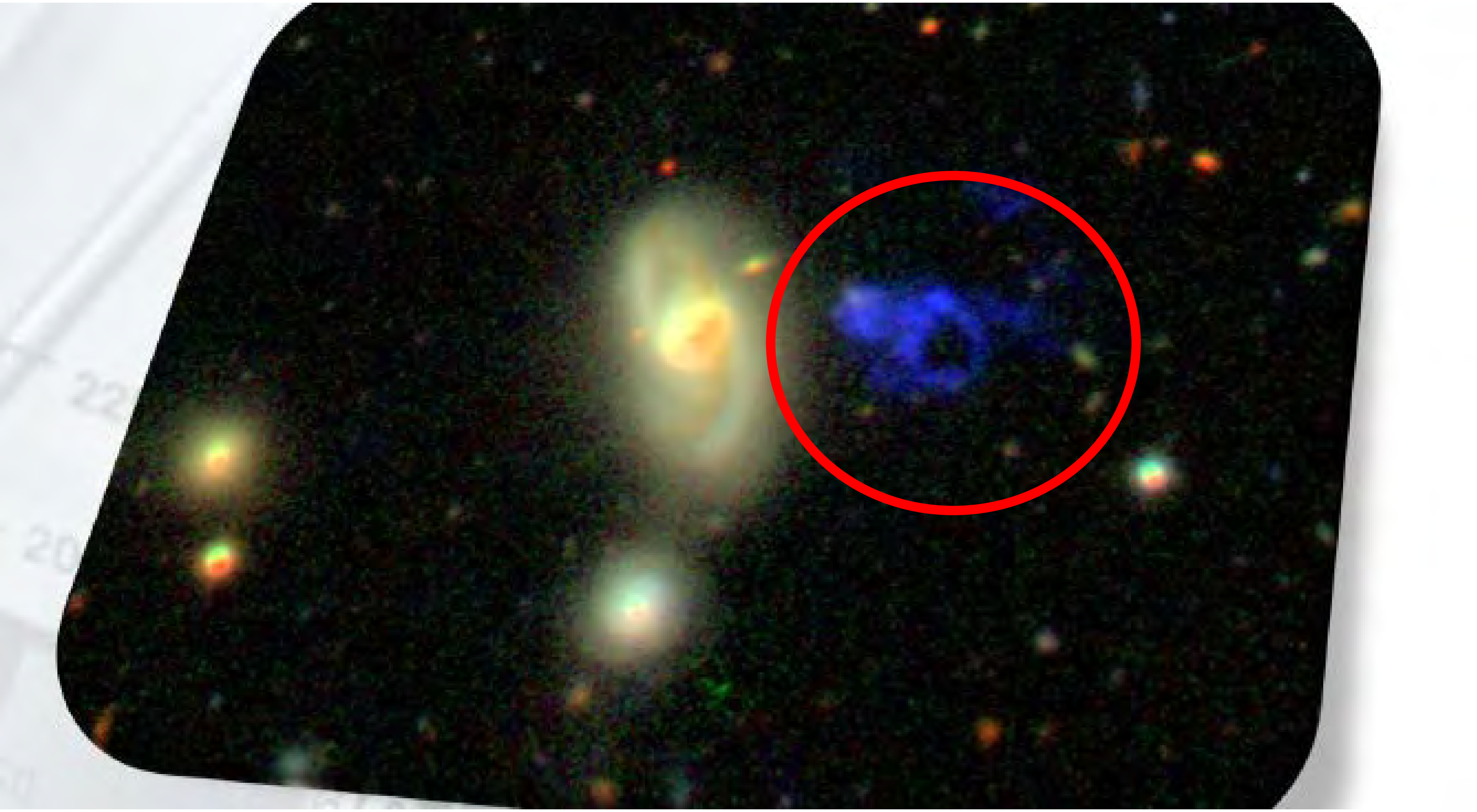
**Info Links**
About Astronomy
About the SDSS
About the SkyServer
SDSS Data Release 6
SDSS Project Website
Open SkyQuery
Images of RC3 Galaxies

**Help**
Getting Started
FAQ
How To
Glossary
Schema Browser
Sample SQL Queries
Details of SDSS Data

Powered by Microsoft

Site Traffic
Privacy Policy

# Hanny van Arkle’s Voorwerp



NATIONAL E-SCIENCE INFRASTRUCTURE CONSORTIUM

# Human Genomes

- Genomic Data is increasing faster than Computing power
- 1000 Genome project
    - Sequence the DNA of the 1,000 human
- ~50GB/person raw data, 6GB/person processed data
- Required storage: ~100TB
- Sequence read: a few month, DNA Sequencing: a few years
- Take a many years to find “Genetic disorder”

# High Energy Physic

- Presentation by Dr. Burin

# Scientific data Today

- Scientific data is doubling every year, reaching EBs (Exabyte)
  - CERN produce 25PB/year
  - Genomes Data ~5PB++
- Data will never will be at a single location
- A lot of un-processed data
- Computing powers are lacking the data
- Soon we cannot even store for the incoming data stream
- Not scalable, not maintainable…

# Challenges

NATIONAL E-SCIENCE INFRASTRUCTURE CONSORTIUM

- Opportunity to access the scientific data
  - Raw, Derived, and Literature data
- New research methodology
  - A scientific revolution in how discovery takes place
    - A rare and unique opportunity (Higg, Voorwerp)
- IT infrastructure to support large data
- High performance computing for processing large amount of data
- Distributed and Shared resources: Grid and Cloud
- Analysis of Distributed Data through Distributed Computing

# National e-Science Infrastructure Consortium

NATIONAL E-SCIENCE INFRASTRUCTURE CONSORTIUM

- Project to provide national infrastructure for e-Science in Thailand
- Total Budget: ~100Million Baht
- Founding Members: Chula, Suranaree, KMUTT, HAII, and NSTDA
- Run as a consortium
- All members manage with their own budget
- Heads of all institutes are committee

- Shared resource through Grid middleware

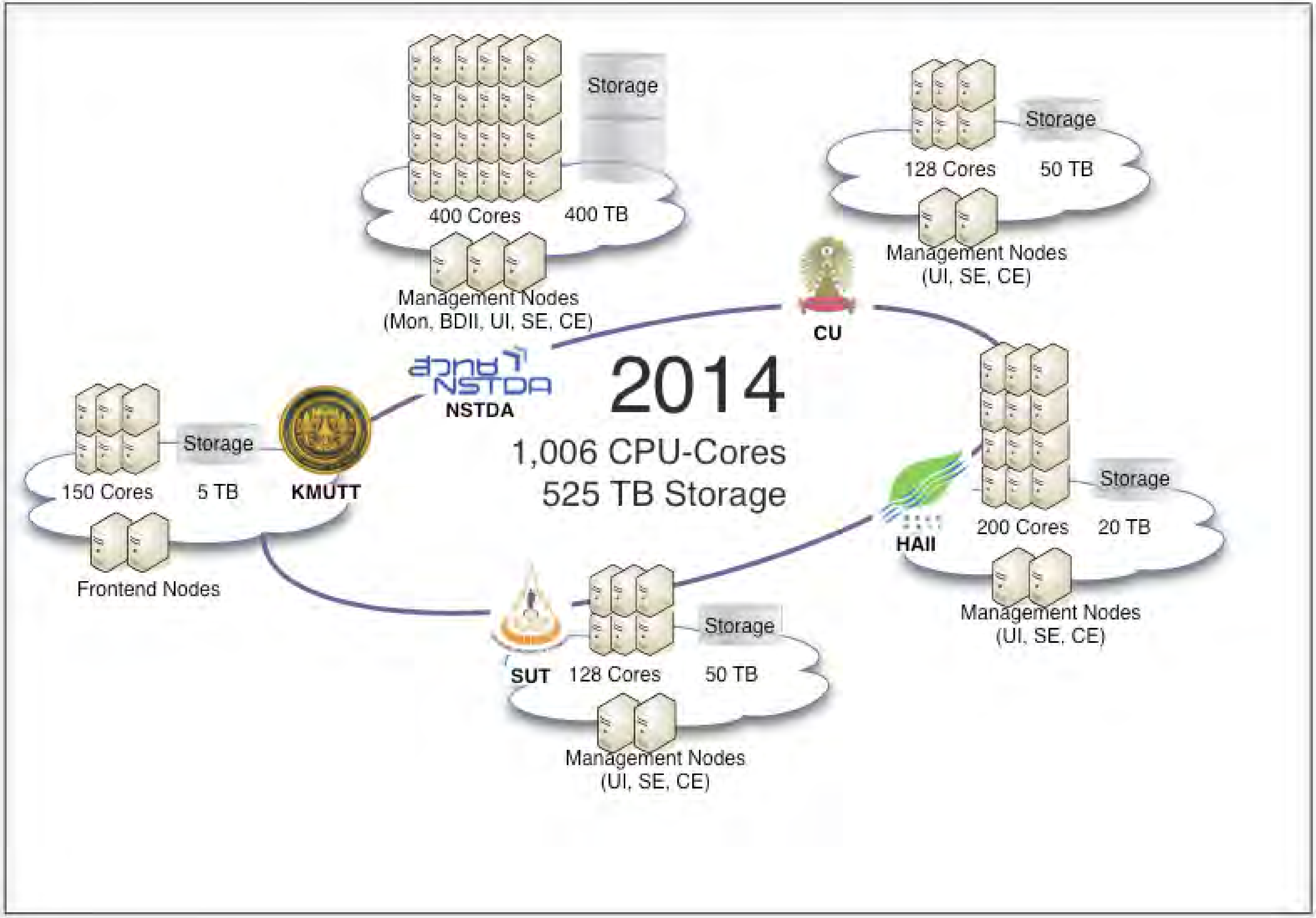
# Current Resources

NATIONAL E-SCIENCE INFRASTRUCTURE CONSORTIUM

- HAII, SUT, and NSTDA Cluster
  - Current total computing cores: 360
  - Total Storage: 350TB

# Resource Roadmap

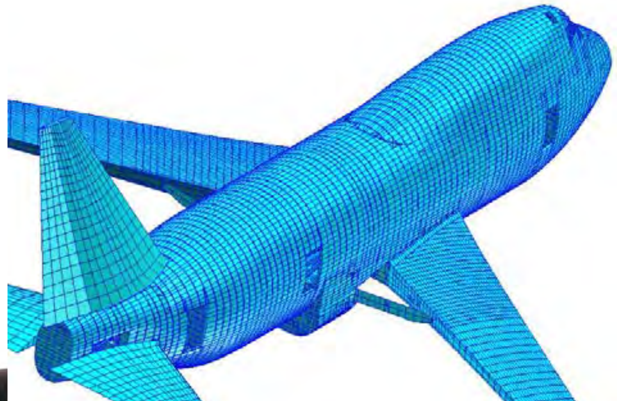
NATIONAL E-SCIENCE INFRASTRUCTURE CONSORTIUM



# Research Area

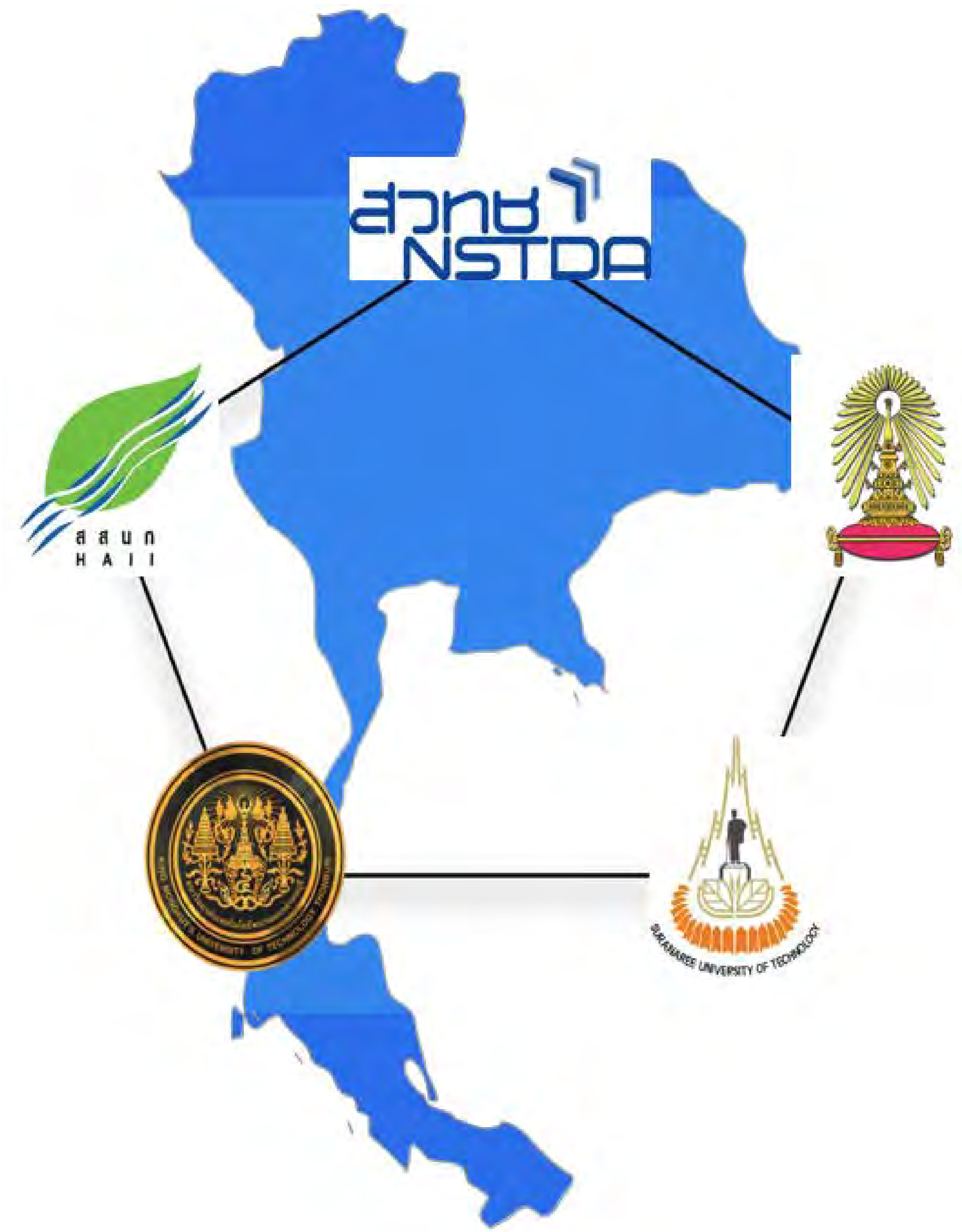
NATIONAL E-SCIENCE INFRASTRUCTURE CONSORTIUM

- High eneygy physics
- Climate change
- Water and energy resources, and, environment
- Computational science and engineering
- Computer science and engineering



Thailand 2011 Flood
oto taken Sep 14, 11: Surajate B.Aroonnet
Photo taken Sep 14, 11: Surajate B.Aroonnet
oto: thairath.co.th
Photo: Channel 3

National e-Science
Infrastructure Consortium



# Thank You