NC STATE UNIVERSITY

Workflow Management Support for Uncertainly Analysis

or

Uncertainty Analytics in the Cloud

Mladen A. Vouk

Professor and Head of Computer Science, and
Associate Vice-Provost for Information Technology

North Carolina State University, Raleigh, NC

At

Workshop on Workshop on Verification, Validation and Uncertainty Analysis in High-Performance Computing (VVUHPC-10)

Supercomputing 2010, November 14, 20101

About the Workshop

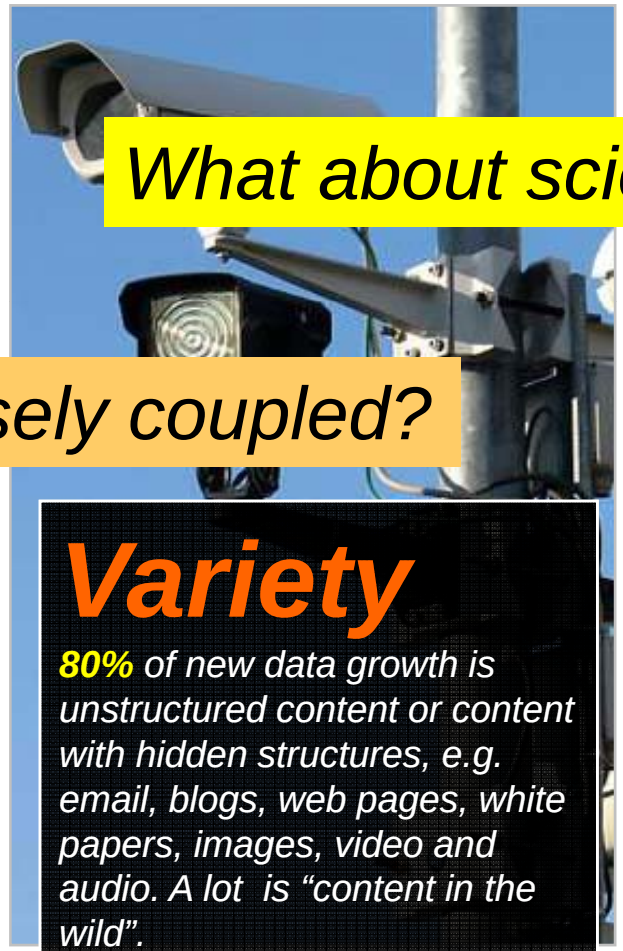
- **Verification, Validation and Uncertainty Analysis in High-Performance Computing**
- *Organizers: Laura L. Pullum (Oak Ridge National Laboratory), Robert M. Patton (Oak Ridge National Laboratory), Thomas E. Potok (Oak Ridge National Laboratory)*
- High-performance computing applications (HPC) have historically advanced the frontier of software complexity, and next generation HPC environments will increase substantially further. The nature of HPC introduces verification, validation and uncertainty analysis (VV&U) challenges that are perhaps unique to the field. Many HPC applications are simulation-oriented which further exacerbates the difficulties by introducing additional validation requirements and possibilities for uncertainty in the results. Unfortunately, HPC application software VV&U do not have a strong tradition since most of the work related to this area has been heavily focused on tolerance to faults due to hardware failure. This workshop will provide a forum for evaluating, sharing, and creating ideas for validation, verification, and uncertainty analysis of HPC applications. For more information go to http://aser.ornl.gov/events/sc2010

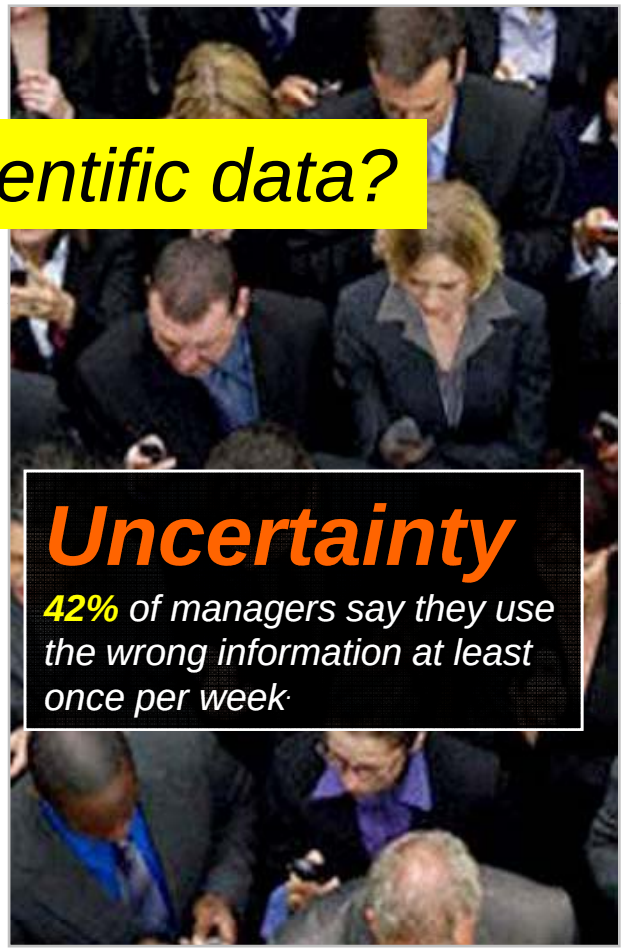
About the Talk

- ***Workflow Management Support for Uncertainly Analysis***
- *Speaker: Mladen A. Vouk (NC State University)*
- This talk discusses functionalities of the DOE SDMC FIESTA scientific workflow support framework (Kepler, Dashboard, Provenance, etc.) that are relevant in the context of uncertainty analyses for both tightly and loosely couple problems. It also presents a cloud-based solution for housing FIESTA (**Framework for Integrated SDM Technologies for Applications)** workflows. NC State's Cloud Computing technology - Apache VCL, service-oriented cloud technology is eminently suitable for construction of sub-clouds, especially analytics sub-clouds. It currently supports Scientific Workflow engine such as Kepler, High-Performance Computing, ample storage, and a range of individual analytics engines and technologies - such as SAS-based, R-based, SPSS- and Cognos-based analytics software, Running uncertainty analyses in this domain is natural.

~ 1 zettabyte on Internet; Challenge: collection, collation, finding, …?

We are experiencing an information explosion





What about scientific data?

Tightly or loosely coupled?

Volume

*Every day, at least **15 petabytes** of new information are being generated. (in 2010 codified information base was expected to double every 11 hours)*

Variety

***80%** of new data growth is unstructured content or content with hidden structures, e.g. email, blogs, web pages, white papers, images, video and audio. A lot is "content in the wild".*

Uncertainty

***42%** of managers say they use the wrong information at least once per week.*

Original slide is courtesy, Dr. Kouri, NCSU BioSci Initiaive, Raleigh, Oct 2010

http://www.er.doe.gov/ascr/Misc/GrandChallenges.html

DOE ASCR Grand Challenges

- Over two years, 9 reports covering: Biology, Climate, Fusion, High-Energy (Quantum) Physics, Security, Nuclear Physics, Nuclear Energy, Modeling and Simulation for Environment, Exascale Workshop
- Grand view (1,327 pages – all 9 reports)
  - Comput(*): 5,598 references
  - Data(*) : 2747 references
  - **Analys**(*) : 938 + **Analytic**(*) : 95 references
  - **<u>Uncertain(*) : 762 references</u>**
  - Secur(*) : 582 references (371 from the security report)
  - Reliab(*) : 220, Fault-toleran (*)
  - I/O : 142 references (input: 220, output 198 references)
  - Workflow(*) : 116 references
  - Cloud(*): 90 references, but only about 5 or so refer to cloud computing
  - GPU(*) : 24

Strategic

Analytics

- Essential in scientific discovery and in everyday life.
  - Data analytics underlie decision support in critical information technology-based or -assisted systems such as health care, power grids, transportation systems, military logistics and tactical systems, intelligence gathering systems, air traffic control, etc.
  - Advanced analytics is also essential for modern financial, marketing and sales decision making.
  - Complexity of associated algorithm and solutions is growing, and the required analysis response times are getting shorter.
- This indicates a need for a very flexible, dynamic and mobile computational platform that minimizes information technology overhead and maximizes end-user effectiveness in the end-user domain.

Strategic

Analytics of Uncertainty

- "Lies, damned lies and analytics"
  - Uncertainty underlies almost all computational, analysis, searching, pattern matching, specification, experimental, etc., activities.
  - How to recognize it, express it, and possibly quantify it is a key element of risk-analysis.
  - Uncertainty analysis is receiving an increasing amount of interest as we move to faster computers, more data, and a greater reliance of all decision making on data analytics, simulations and computationally supported experiments.
- Risk = P(unfavorable event) * loss
- Error, upper/lower bound, range, conditional probability, sensitivity analysis, pattern match, social behavior variance, etc., etc.

Analytics Cloud (2)

- A flexible distributed computing model (cloud) works for many uncertainty computations and that workflow support is needed to manage that.
- Because
  - Computational environment needs to be sensitive to data location, domain and analytics, as well as security, privacy and preferences issues.
  - It needs to optimize workloads/workflows according to identified "bottlenecks" – from data-models and real physical issues (e.g., data movement rates, to security concerns, to perceived problems based on user preferences, to application characteristics, etc.).
  - Configuration needs to be flexible ("elastic") - details depend on whether differentiated (can be re-configured at will by end-user) or undifferentiated (can only be used "as is") resources are used, and whether analytics moves to the data, or it is possible to move data to analytics engines.

Supporting Uncertainty Analysis

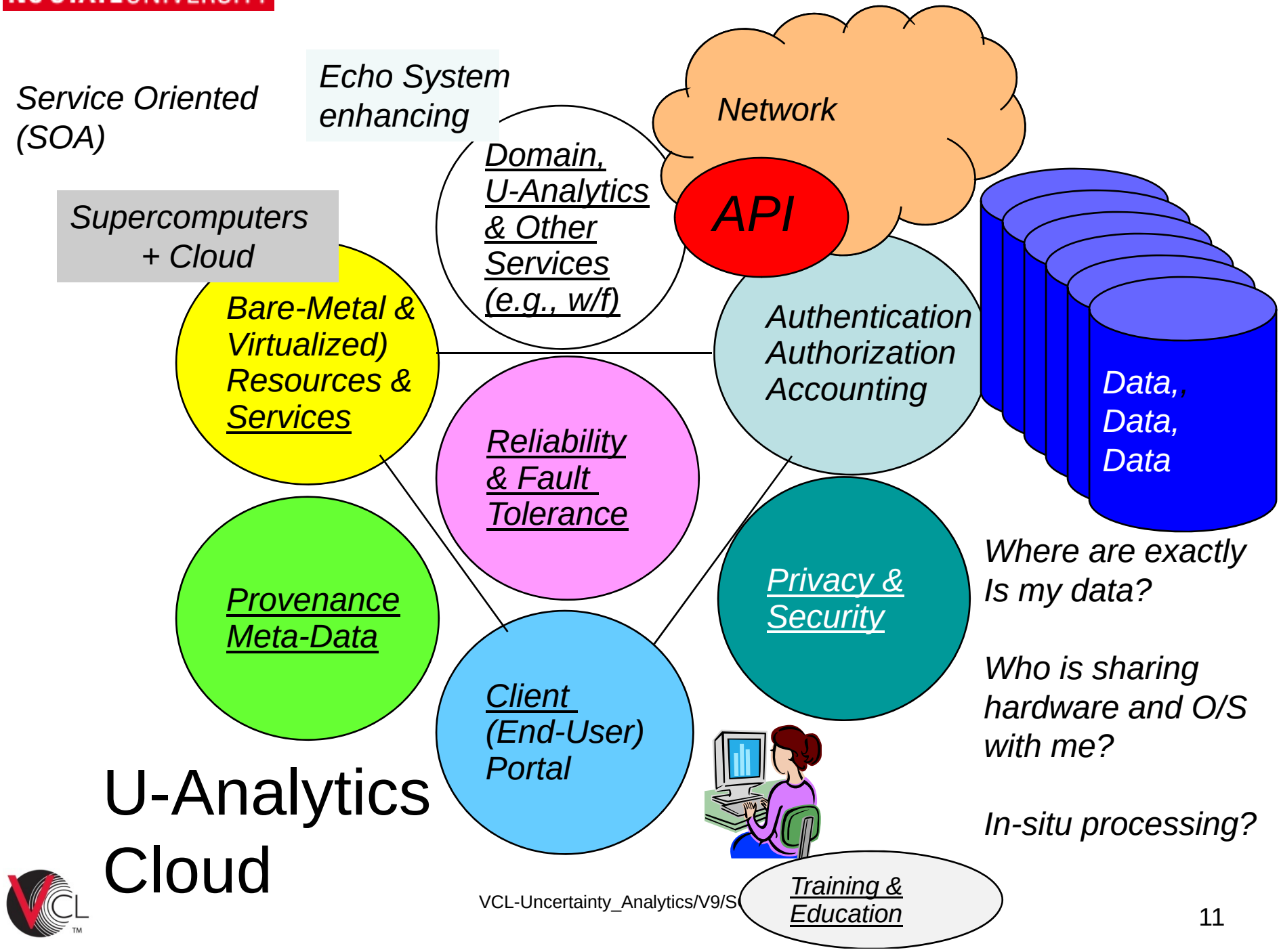
- Facilities
  - HPC shared and distribtued memory modeling, hybrid models
  - Ability to launch a number of what-if simulations, repetitive tasks
  - **Workflow support**
  - Storage management
  - Sub-cloud generation (**Cloud Computing**)
  - Version management
- There are many uncertainty analytics needs
  - Text analytics and uncertainty software
  - Statistically-based uncertainty software
  - Reliability of model prediction, comparisons
  - Risk modeling, security assessment, noise recognition
  - Tools for assessing uncertainty in unstructured, qualitative, quantitative, and “in the wild” data,
  - Other …

Strategic

U-Analytics and Security

- *Secure Analytics (how to do analytics securely)*
  - Information technology related security for critical systems often focuses on the use of intrusion detection, firewalls, and encryption to protect the infrastructure and the endpoints of data transmission and exchange.
  - This is far from enough. Ensuring that analytics is done in a secure and privacy-preserving manner is paramount for the soundness of the recommendations that come from the analytics, and for preserving the privacy and reputations of individuals and organizations whose data is analyzed.
  - In addition to traditional data security, analysis also has to focus on the analytics engines and their mode of use, as well as on the training end-users have received.
- *Security Analytics (how to assure security through analytics)*
  - Furthermore, analytics can be used to check on the security of software-based systems and insure that they are secure and privacy-preserving.
- Both cases require considerable knowledge of the domain in which the transactions, data collection, and data processing occur.

NC STATE UNIVERSITY



Where are exactly Is my data?

Who is sharing hardware and O/S with me?

In-situ processing?

U-Analytics Cloud

NC STATE UNIVERSITY

Office of Science U.S. DEPARTMENT OF ENERGY | SDM CENTER | Kepler

Key Partners

How to do it: Workflows + Cloud

Using Department of Energy
Scientific Data Management Center for Enabling Technologies
Framework for Integrated SDM Technologies for Applications (FIESTA)

https://sdm.lbl.gov/twiki/bin/view/SDMCenter/WebHome

Using Open Source Kepler Workflow Technology (https://kepler-project.org)

Using NCSU VCL Cloud Technology (http://vcl.ncsu.edu)

Why Workflows Work?

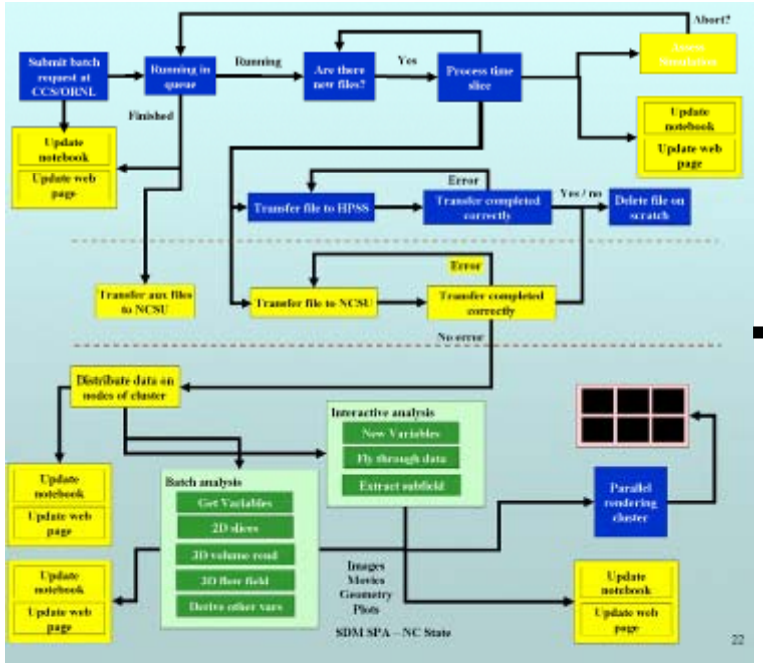
- Uncertainty analysis often requires
  - Repeated data movements
  - Thousands of runs and possibly runs at many different locations (perhaps with different parameters)
  - Job monitoring
  - Post-simulation data transfers
  - Post-processing analytics
  - Frequent repetition of the all of the above after the models, data, or inputs have been changed
  - Other …
- Many workflow support systems exist

Example – Computational Astrophysics

- Dr. Blondin is carrying out research in the field of Circumstellar Gas-Dynamics. The numerical hydrodynamical code VH-1 is used on supercomputers, to study a vast array of objects observed by astronomers both from ground-based observatories and from orbiting satellites.
- The two primary subjects under investigation are interacting binary stars - including normal stars like the Algol binary, and compact object systems like the high mass X-ray binary SMC X-1 - and supernova remnants - from very young, like SNR 1987a, to older remnants like the Cygnus Loop.
- Other astrophysical processes of current interest include radiatively driven winds from hot stars, the interaction of stellar winds with the interstellar medium, the stability of radiative shockwaves, the propagation of jets from young stellar objects, and the formation of globular clusters.

NC STATE UNIVERSITY

The Big Picture: Supporting the Scientist

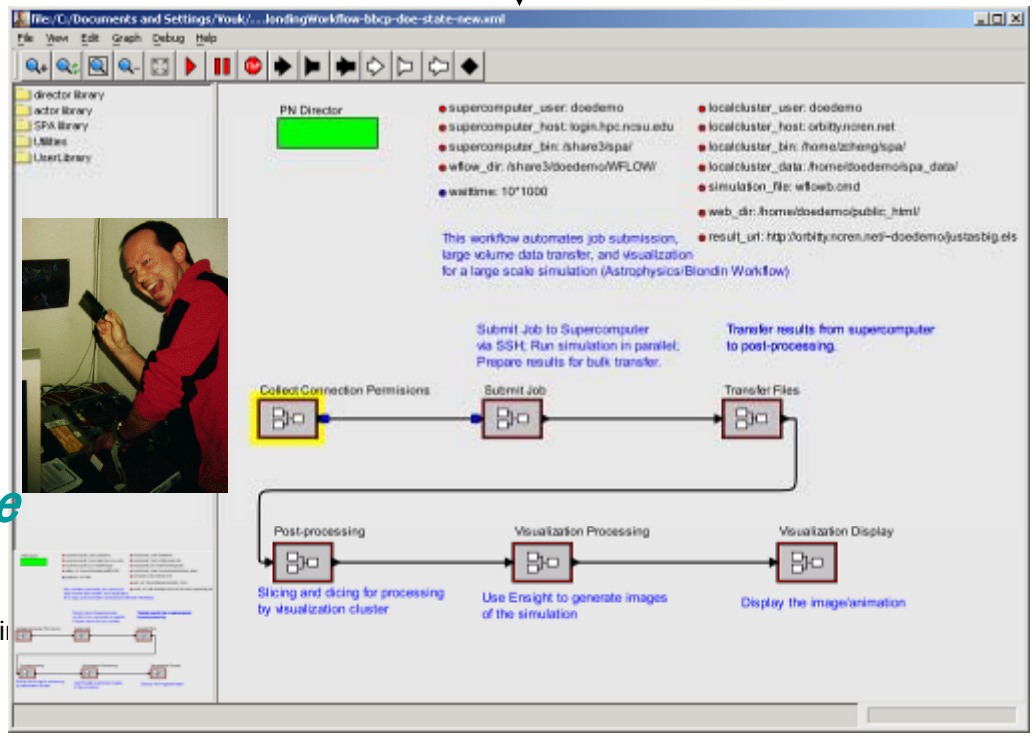


From "Napkin Drawings" …

… to Executable Workflows

Conceptual SWF

Executable SWF



Here:

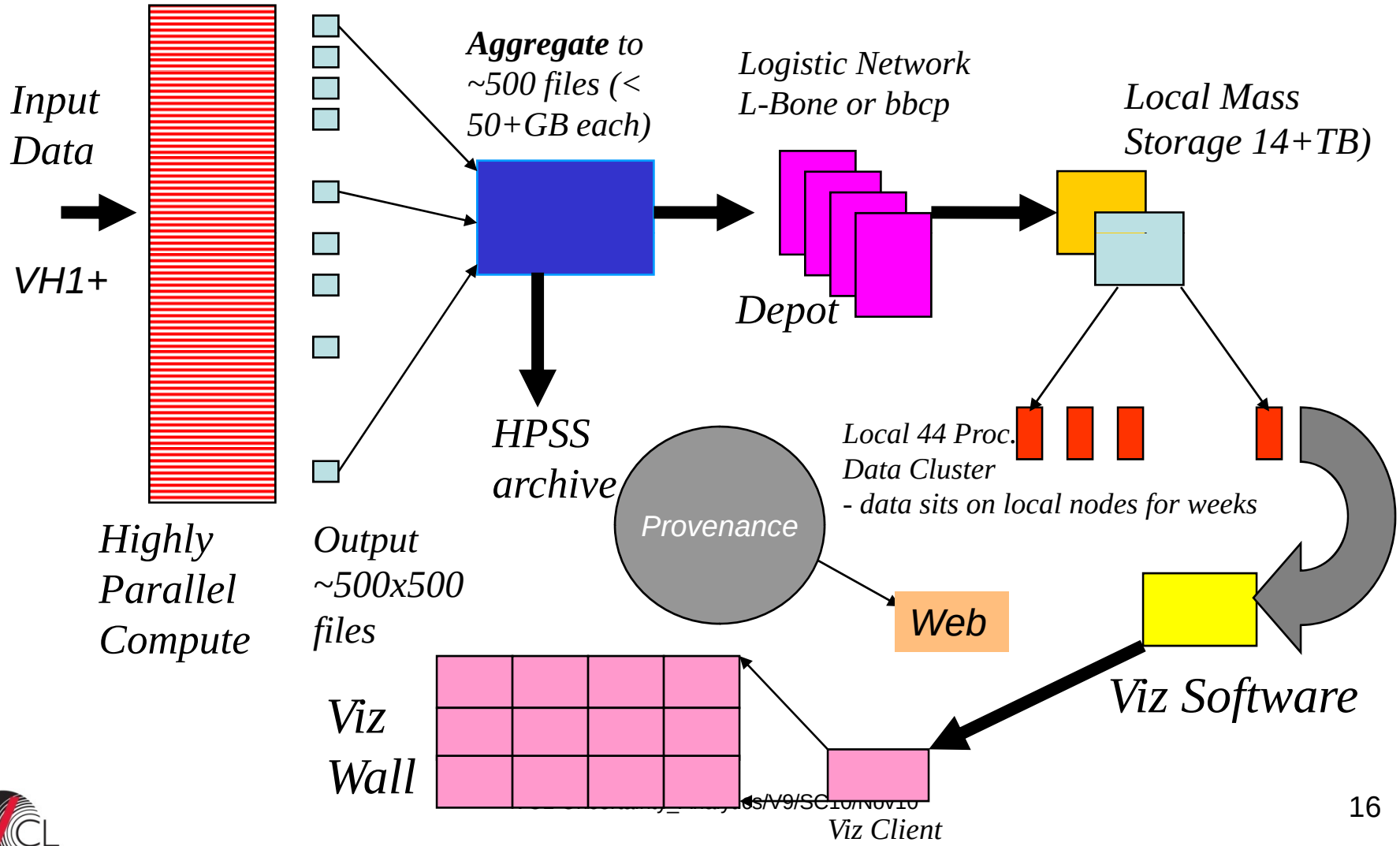
John Blondin, NC State
Astrophysics
Terascale Supernova Initiative
SciDAC, DOE

NC STATE UNIVERSITY

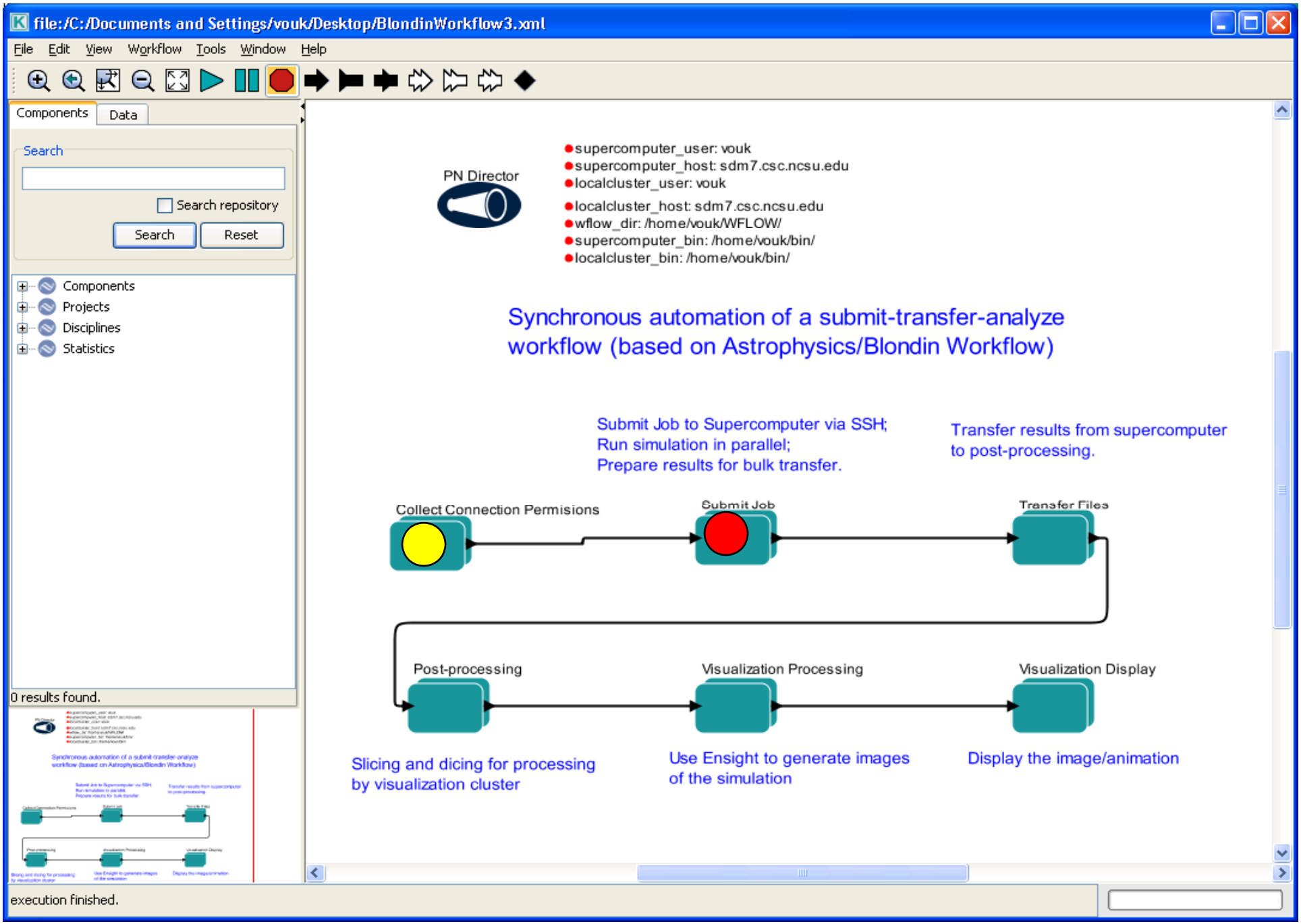
Scientific Workflow Automation (e.g., Astrophysics 2007)

In conjunction with John Blondin, NC State University

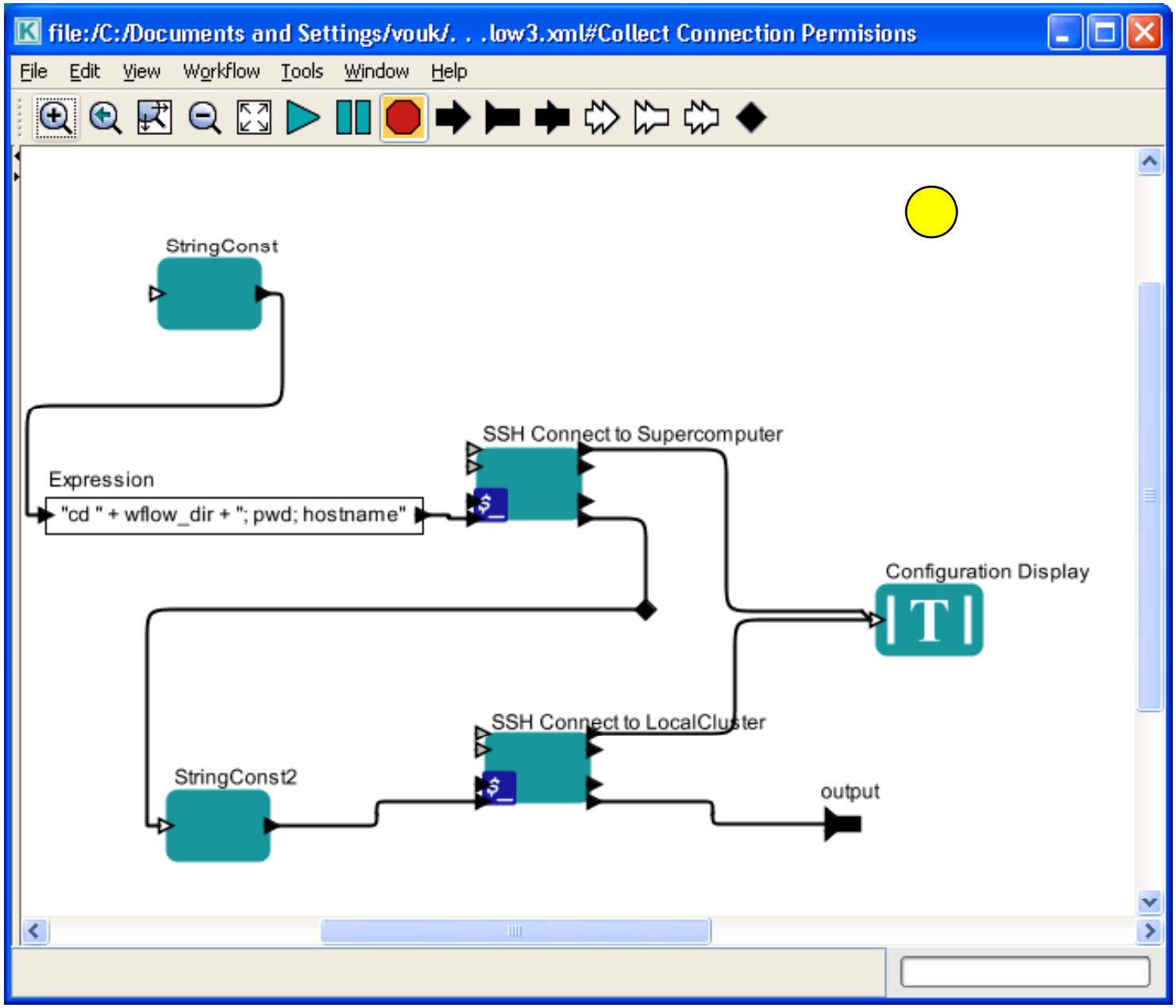
Automate data acquisition, transfer and visualization of a large-scale simulation at ORNL

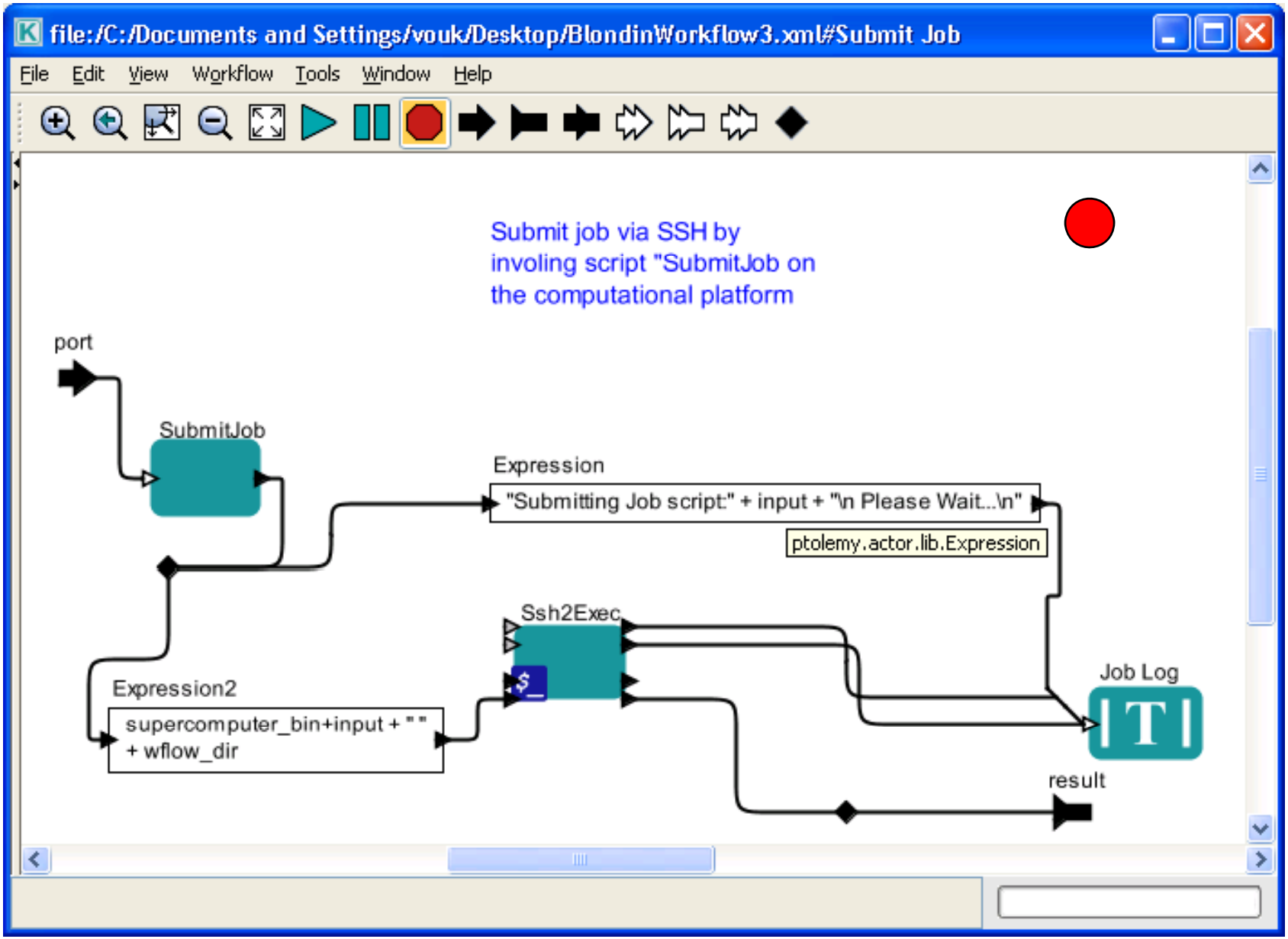


NC STATE UNIVERSITY

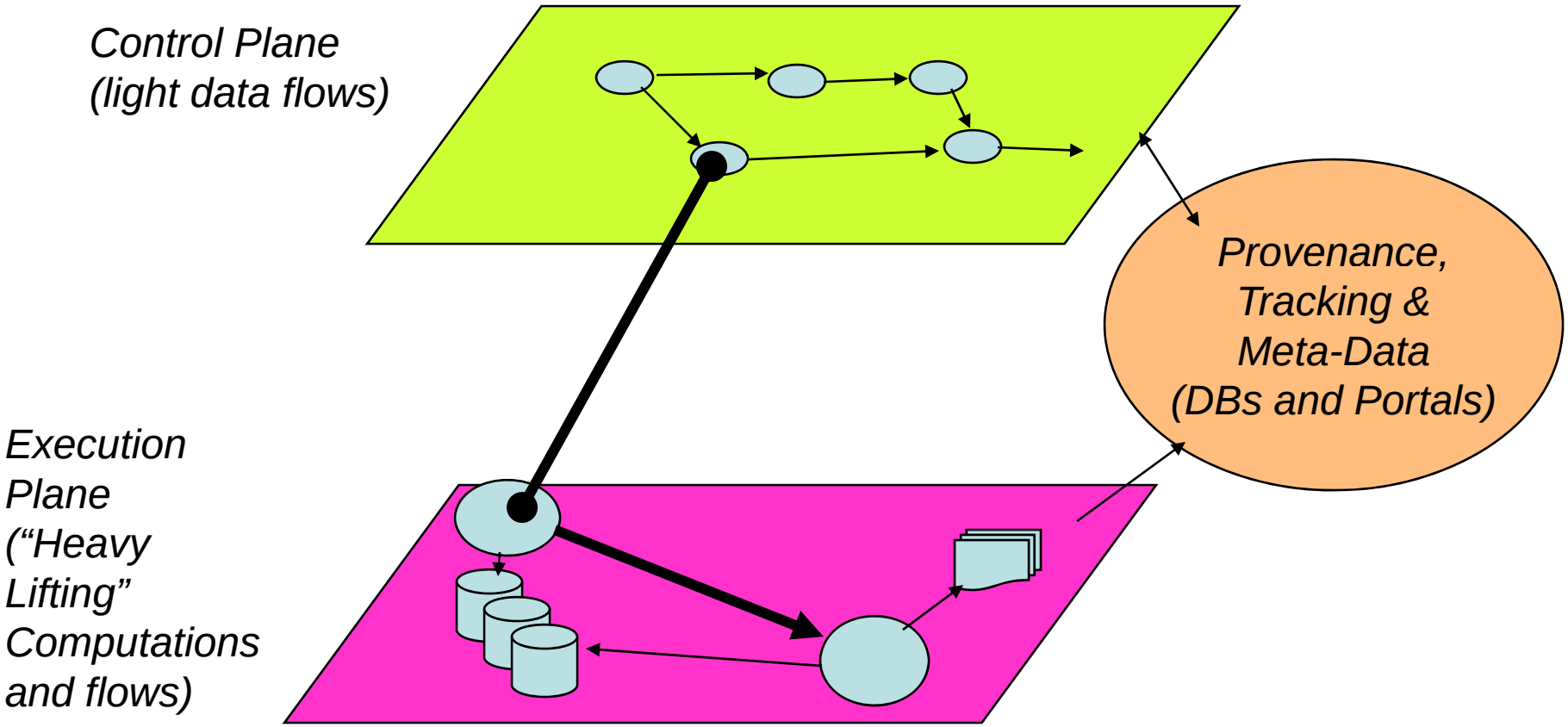


NC STATE UNIVERSITY




file:/C:/Documents and Settings/vouk/Desktop/BlondinWorkflow3.xml#Submit Job
File Edit View Workflow Tools Window Help
Submit job via SSH by
involing script "SubmitJob on
the computational platform
port
SubmitJob
Expression
"Submitting Job script:" + input + "\n Please Wait...\n"
ptolemy.actor.lib.Expression
Ssh2Exec
Expression2
supercomputer_bin+input + " "
+ wflow_dir
Job Log
result

An Extended View of an Actor



Synchronous or Asynchronous?

NC STATE UNIVERSITY

Actor in a Broader Sense

In

jar:file:/C:/Kepler/build/kepler.jar!/org/sdm/spa/Ssh2Exec.java

File Tools Help

```
////////////////////////////////////////////////////////////////////////////////
//// Ssh2Exec
/**
   <p>Connects to a remote host using Ssh2 protocol.

   </p><p>Error conditions this actor must respond robustly to:
   <ul>
   <li>Wrong identity file given.</li>
   <li>Host unreachable.</li>
   <li>Login unsuccessful.</li>
   <li>Session dies prematurely.</li>
   </ul>

   </p><p>This actor will keep the session open until it receives a different
```

Out

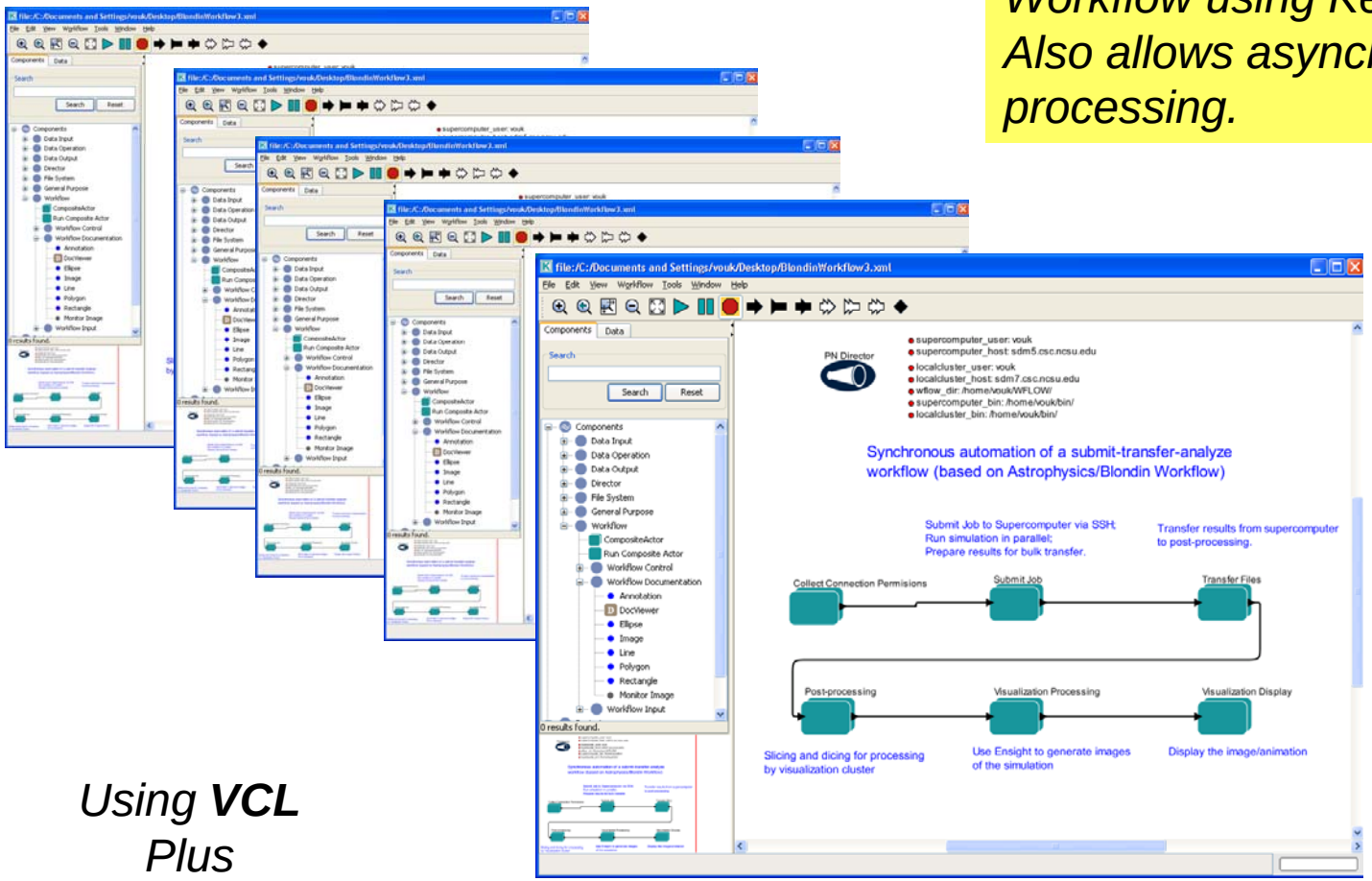
Network/"Cloud"

```
Bsub < code_run
------------ where code_run is a script --------------
code_run
#! /bin/csh
source /usr/local/lsf/conf/cshrc.lsf
#BSUB -W 5
#BSUB -n 100
mpiexec ./code
#BSUB -o /share/vouk/WFLOW/code.out.%J
#BSUB -e /share/vouk/WFLOW/code.err.%J
#BSUB -J codevouk
-------------------------
```

NC STATE UNIVERSITY

Parallel & Virtual

Multiple instances of the Workflow using Kepler. Also allows asynchronous processing.



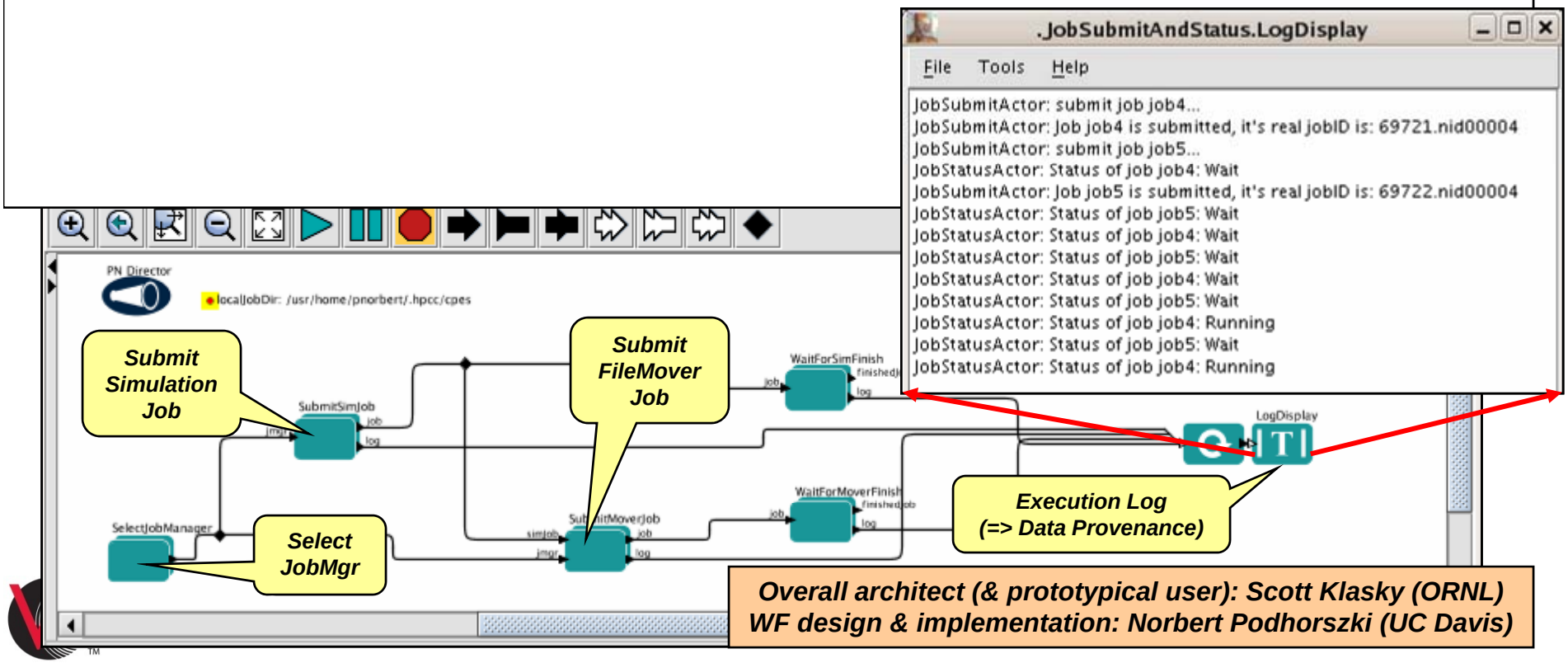
*Using **VCL***
Plus

Windows, Mac and Linux Clients

NC STATE UNIVERSITY

CPES Fusion Simulation Workflow

- **Fusion Simulation Codes**: (a) GTC; (b) XGC with M3D
- **GOAL**:
  - automate remote simulation **job submission**
  - continuous **file movement** to secondary **analysis cluster** for dynamic visualization & simulation control
  - … with **runtime-configurable observables**



Overall architect (& prototypical user): Scott Klasky (ORNL)
WF design & implementation: Norbert Podhorszki (UC Davis)

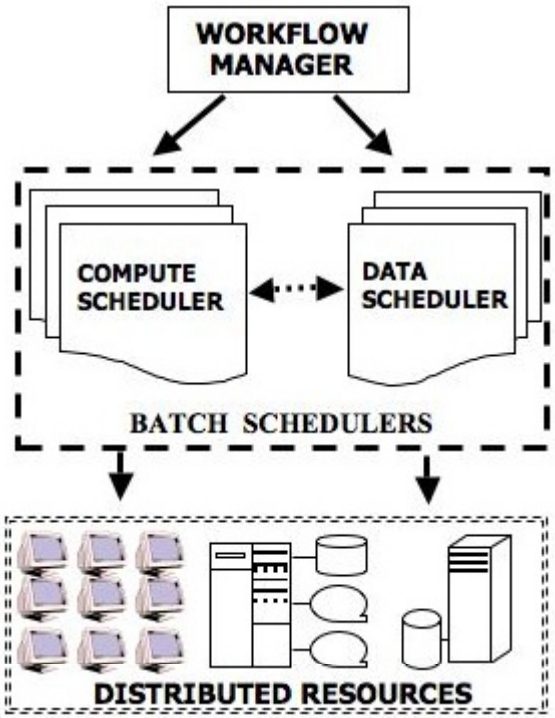
https://kepler-project.org/users/projects-using-kepler

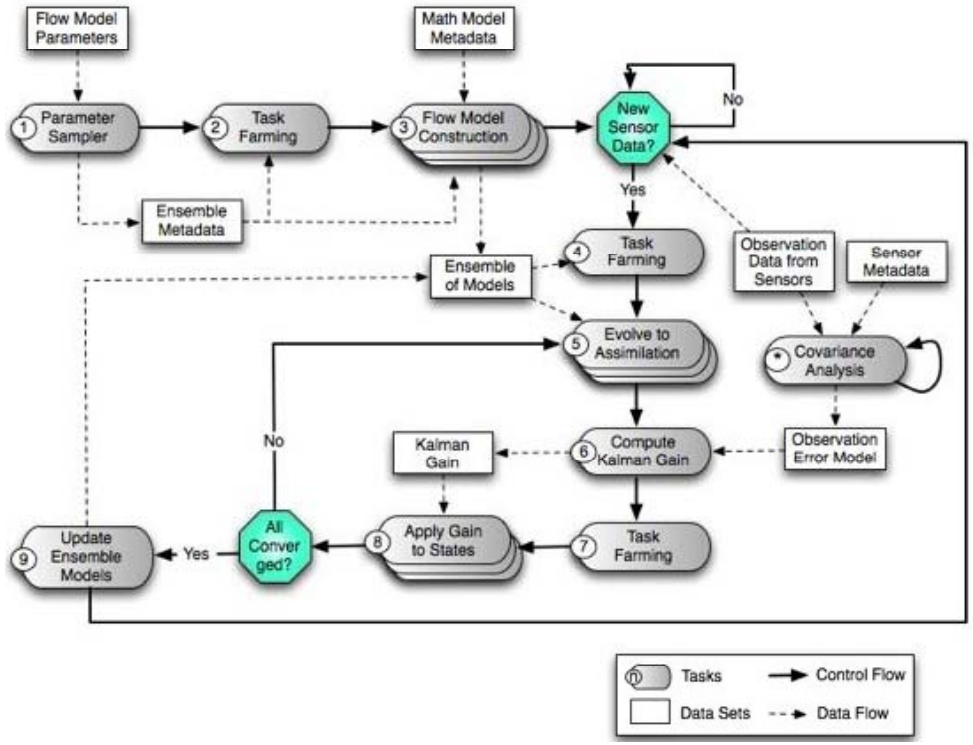
Other Projects Using Kelper

- Dozens of projects are using Kepler to solve a variety of technical challenges. For example,
- Real time sensor data analytics

NC STATE UNIVERSITY

A Grid-enabled Workflow System for Reservoir Uncertainty Analysis*





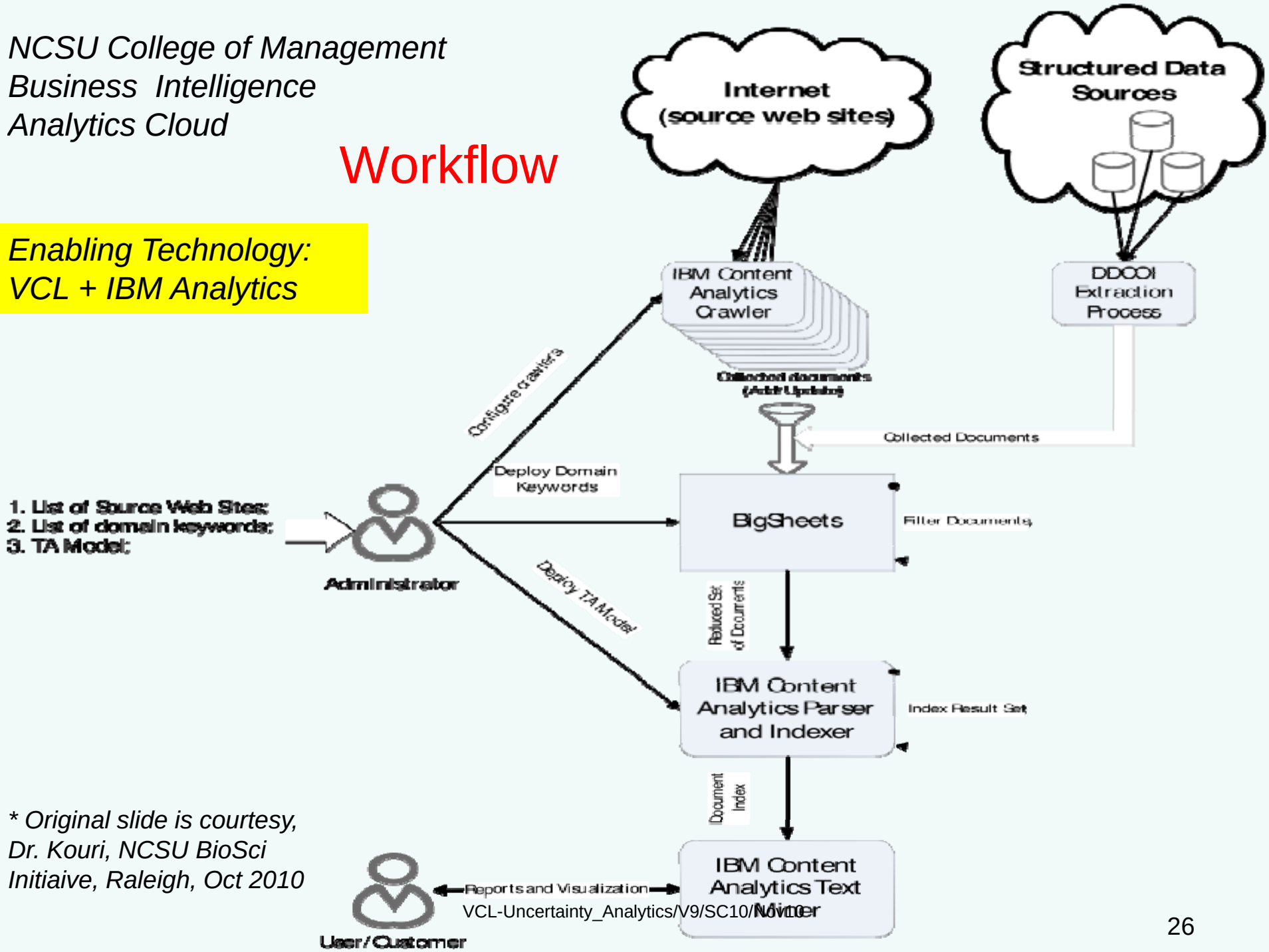
Enabling Technologies: Grid and Workflow support via Condor-G, DAGMan, and Stork.

(*) From the paper by Emrah Ceyhan, Gabrielle Allen, Christopher White and Tevfik Kosar, Center for Computation & Technology, Louisiana State University, Baton Rouge, LA 70803, USA, Proceedings of CLADE'08 (in conjunction with HPDC'08), p 45-52, Boston, MA, (2008).

NCSU College of Management
Business Intelligence
Analytics Cloud

Workflow

Enabling Technology:
VCL + IBM Analytics

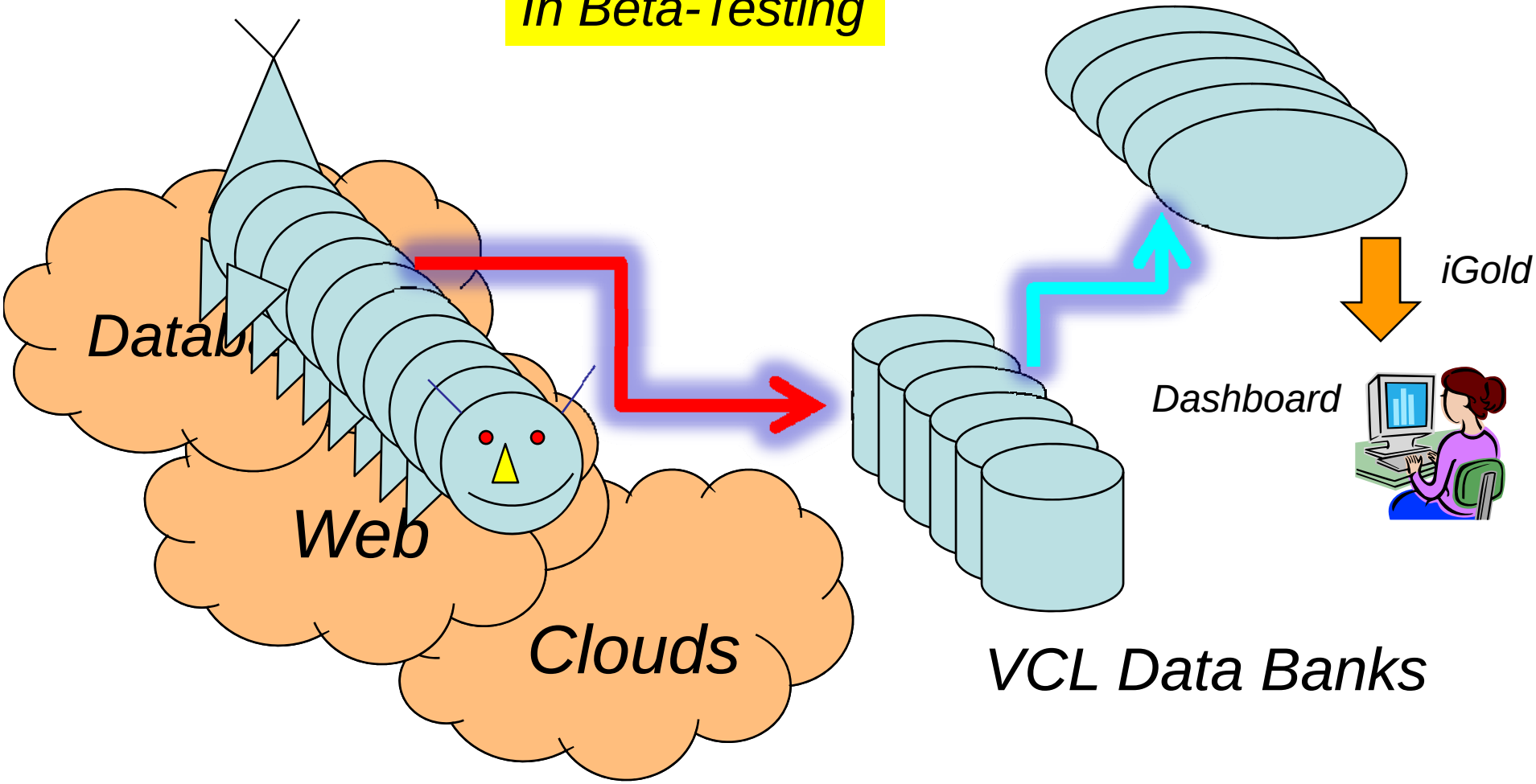


** Original slide is courtesy,*
Dr. Kouri, NCSU BioSci
Initiaive, Raleigh, Oct 2010

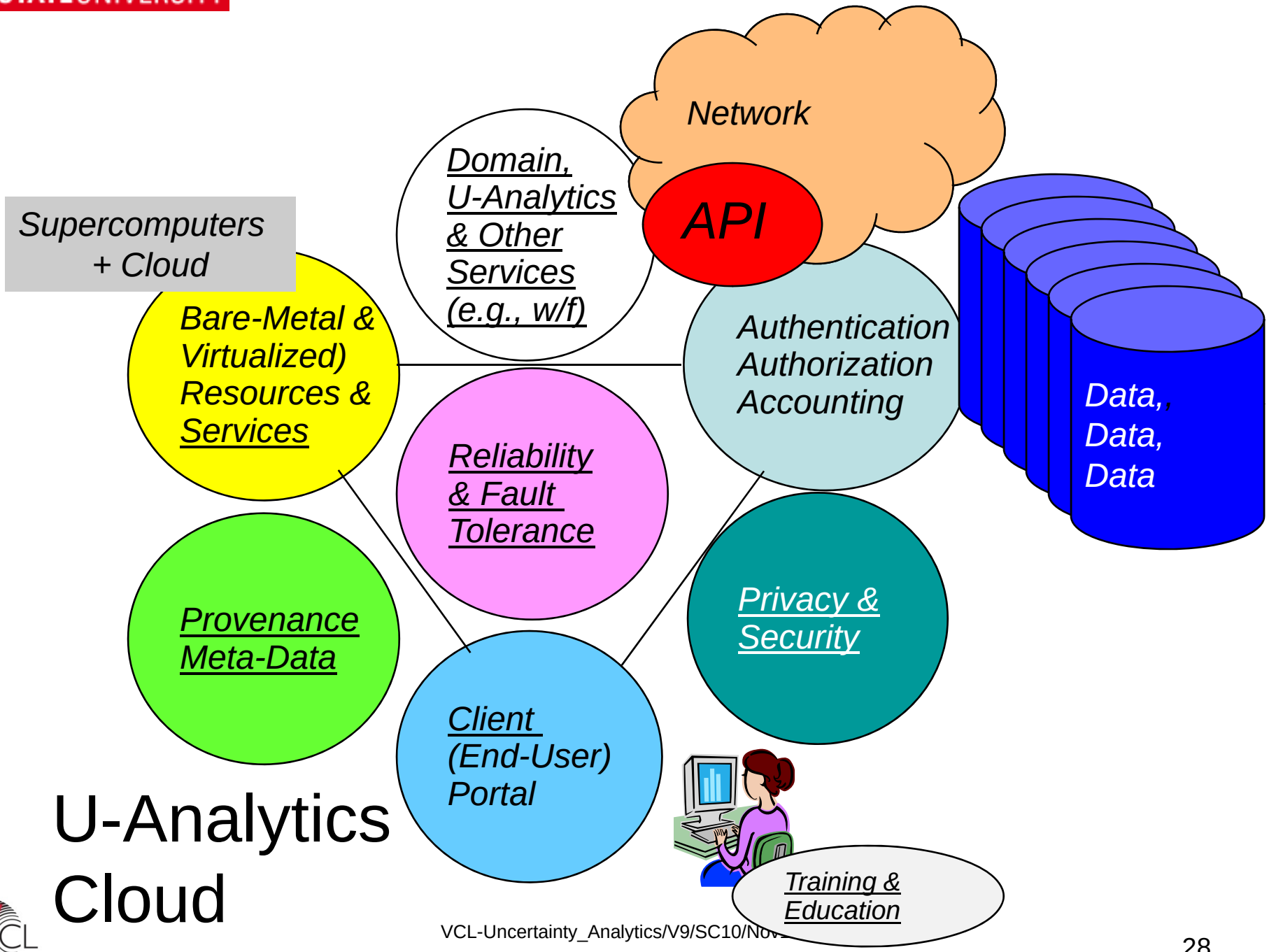
NC STATE UNIVERSITY

Analytics Cloud

In Beta-Testing

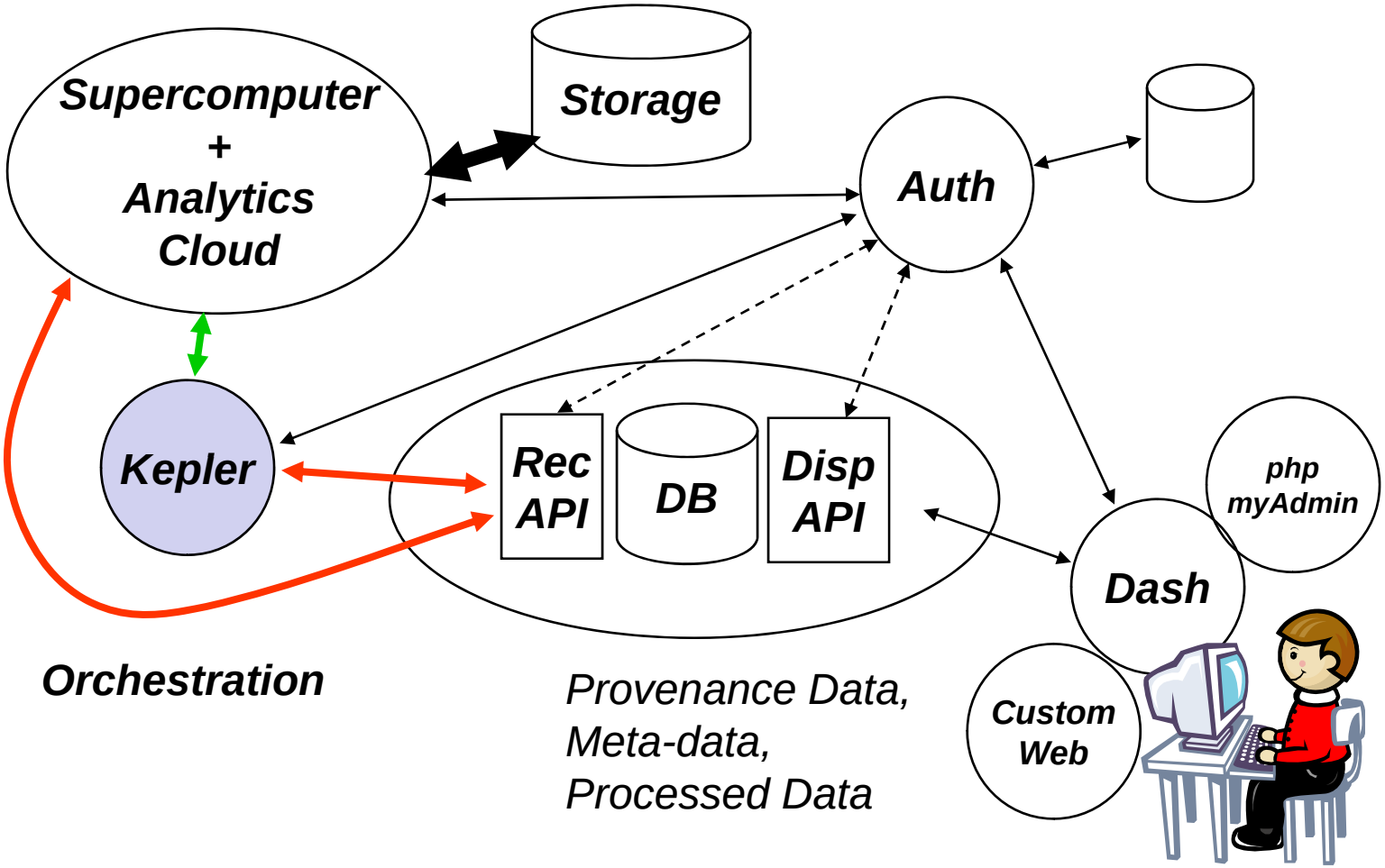


VCL Centipede Crawler Images



U-Analytics Cloud

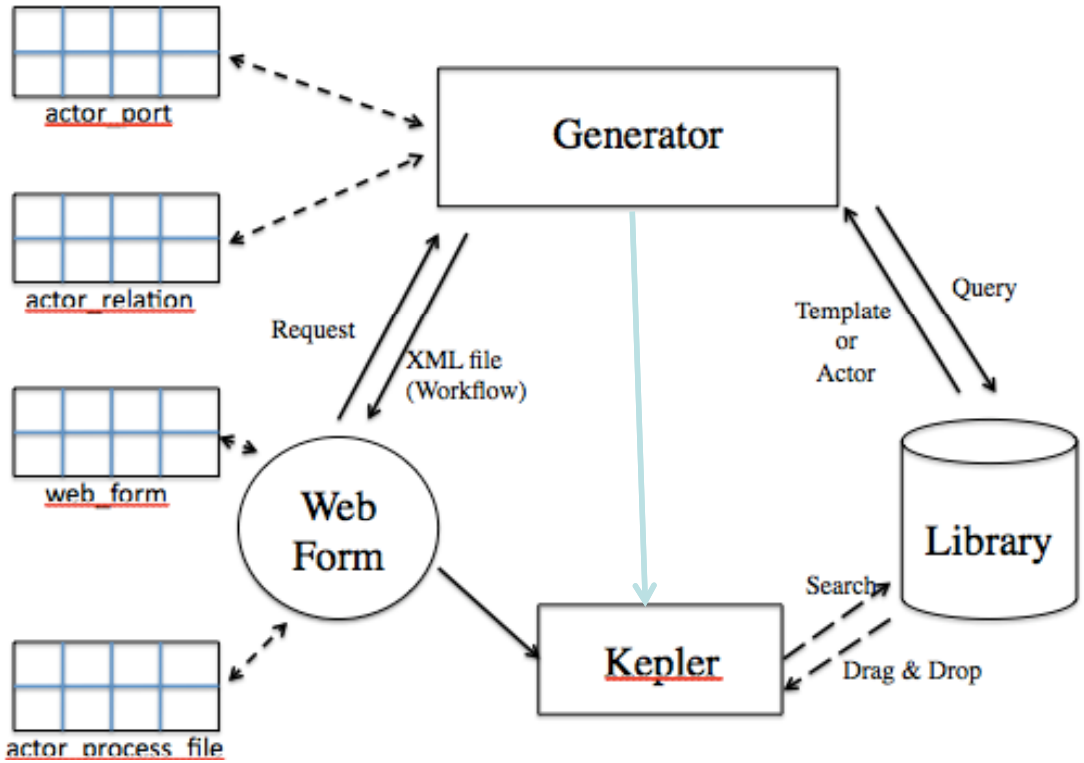
The FIESTA Framework for workflow management



NC STATE UNIVERSITY

Workflow Generator Wizard

- **Web Form**
  - User fills in the options in the webform and clicks on the submit button.
- **Library**
  - Contains the actors and the templates
- **Generator**
  - Extracts the information from the web form and uses the library to build the workflow.
  - It uses Kepler to view and execute the workflow.



Architecture

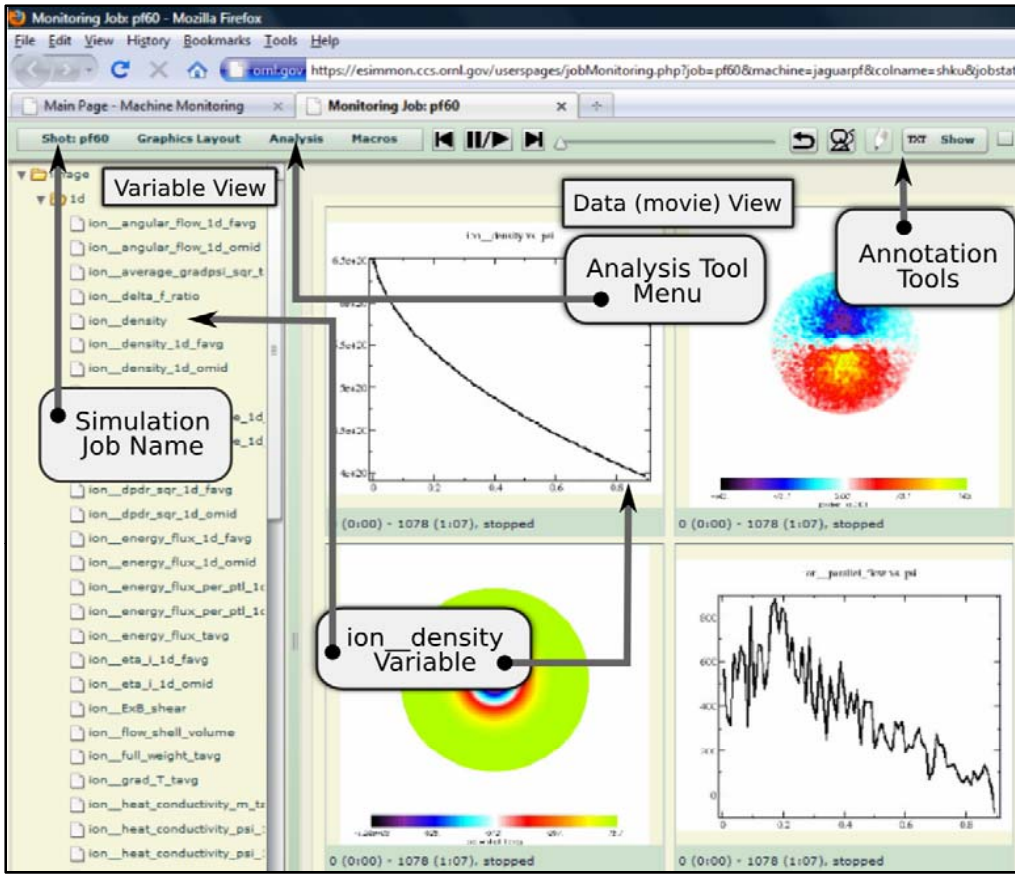
Level 2 Support

NC STATE UNIVERSITY

Dashboard

Some Characteristics

- ***Functionality***: e.g., At a glance view the status and health of a simulation. Scientists can monitor running simulation or explore and manipulate data from past simulations from any web browser.
- ***Ease of use:*** e.g., Doesn't require the end user to be an expert on File Systems or Visualization tools
- ***Collaboration:*** e.g., Alow scientists to collaborate and focus just on the science. The eSimMon dashboard organizes users and collaborators simulation runs (shots).
- ***Annotation:*** e.g., Users can annotate shots and search through several shots archived.
- ***Data Reduction and Abstraction***: raw data into
- ***Analysis:*** Provides several types of analysis tools
- Etc.



Portable Dashboard

NC STATE UNIVERSITY

IBM

Key Partner

http://vcl.ncsu.edu

How to do U-Analytics in a Cloud

Virtual Computing Laboratory Technology is Open Source

http://incubator.apache.org/projects/vcl.html

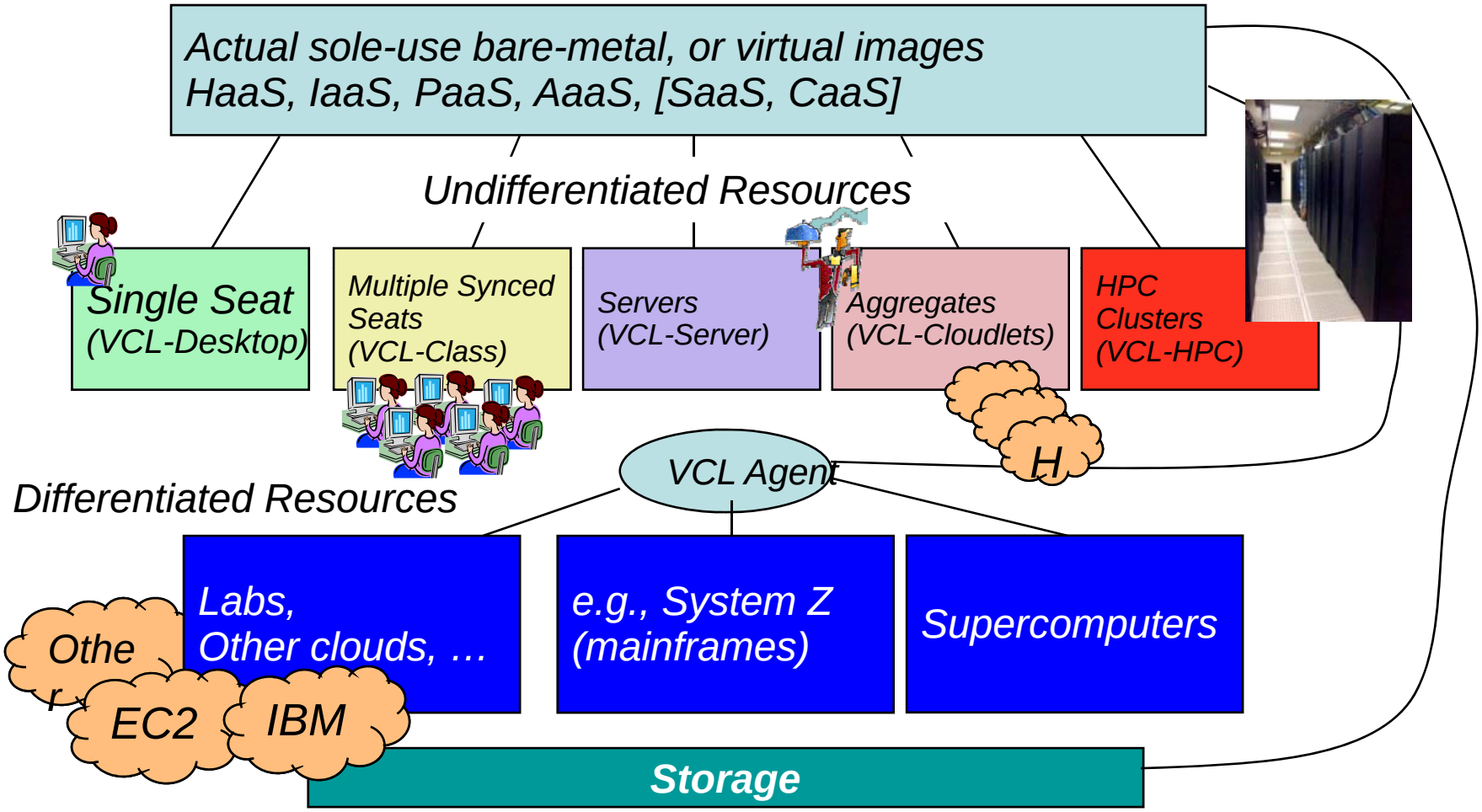
Current NC State University VCL installation **– Private Cloud**:

- 2000+ blades
- About 800+ in General mode, about the same in HPC mode, and several hundred in various test-beds
- open to 40,000+ NCSU students and faculty, different pilots and partner accounts, through Shibboleth all UNC System campuses have access to VCL (cca 250,000 students).
- 250,000+ reservations per year + 10.5 million CPU hours
- Linux, Windows environments

NC STATE UNIVERSITY

Current VCL Services

Reservation Times:
From 30 min to open-ended
Load Times:
From 30 sec to 20 min
(service dependent, 80+% in less than 2 min)



Accommodates Different Models

- Range
  - Tightly coupled (often numerical models running over grids)
  - Loosely coupled (often embarrassingly parallel)
  - Movable Data
  - Immovable Data
- Adaptability, Hybrid models (part tightly, part loosely coupled)
- Analytics
  - Text Analytics
  - “Classical Analytics” – Statistics
  - Visual Analytics
  - Other

Home | Virtual Computing Lab (VCL) - Windows Internet Explorer

http://vcl.ncsu.edu/

File Edit View Favorites Tools Help

Convert Select Contribute Edit Post to Blog

Favorites Home | Virtual Computing Lab (VCL)

Page Safety Tools

NC STATE UNIVERSITY | Office of Information Technology | OIT Units | Help Desk | Sysnews | Search OIT

VIRTUAL COMPUTING LAB
powered by Apache VCL

VCL LOGIN

View site with SSL enabled to see login form

Log in with NCSU WRAP
View all authentication options

Home

SEARCH

Home Page

- Reservation System
  - Make a Reservation
  - Current Reservations
- Help & Support
  - How Do I Get Started
  - Access Requirements
  - Frequently Asked Questions
  - Documentation
- The NC Education VCL Consortium
  - Academic Partners
  - Corporate Partners

PROVIDING SOLUTIONS FOR

Remote Access Service

NCSU provides a remote access service that allows you to reserve a computer with a desired set of applications for yourself, and remotely access it over the Internet. This service is powered by the Apache Software Foundation's Virtual Computing Lab (VCL) software.

You can use all your favorite applications such as Matlab, Maple, SAS, Solidworks, and many others. Linux, Solaris and numerous Windows environments are now available to all NC State students and faculty. Instructors that would like to request course access, please see these instructions.

History

Development of The Virtual Computing Lab (VCL) started in 2004 as a joint venture of the College of Engineering (COE) and the Office of Information Technology (OIT) to efficently use hardware investments and to provide remote access to a wide range of advanced compute requirements by both NCSU students, faculty, and researchers.

VCL NEWS

- Congressman Price kicks off NC Bio-Prepared
- NetApp's Support Enables NC State's Next Generation Cloud Computing Research
- GMU Receives 2009 Virginia Governor's Technology Award for use of VCL
- NC State & IBM host 3rd International Conference on the Virtual Computing Initiative - October 22-23, 2009
- Duke Starts VCL Pilot
- VCL part of National Science Foundation Grant

See All News Items

VCL SYSTEM STATUS

VCL blades online: 652

Done | Internet | Protected Mode: On | 100%

VCL: The Virtual Computing Lab - Windows Internet Explorer

https://vcl.ncsu.edu/scheduling/index.php?mode=newRequest

VIRTUAL COMPUTING LAB
powered by Apache VCL

Logged in as vouk from NCSU
Log out

Home » Reservation System

- New Reservation
- Current Reservations
- Block Allocations
- User Preferences
- Manage Groups
- Manage Images
- Manage Schedules
- Manage Computers
- Management Nodes
- View Time Table
- Privileges
- User Lookup
- Virtual Hosts
- Site Maintenance
- Statistics

New Reservation

Please select the environment you want to use from the list:

Access Visio Excel 2003 (WinXP)

Image Description:
Access Visio Excel 2003
putty, wincp
adobe reader 9

When would you like to use the application?
Now
Later: Sunday At 1 45 p.m. (Eastern Time Zone)

Duration: 1 hour
Until (YYYY-MM-DD HH:MM:SS)

Estimated load time: < 1 minute

Create Reservation

General Reservation (VCL-Desktop, VCL-Server)

Short-term reservations (2-3 hours)

Long-term reservations
Need to explicitly manage
- state persistence
- timeout
- backups

Frequently used image load very quickly

Done | Internet | Protected Mode: On | 100%

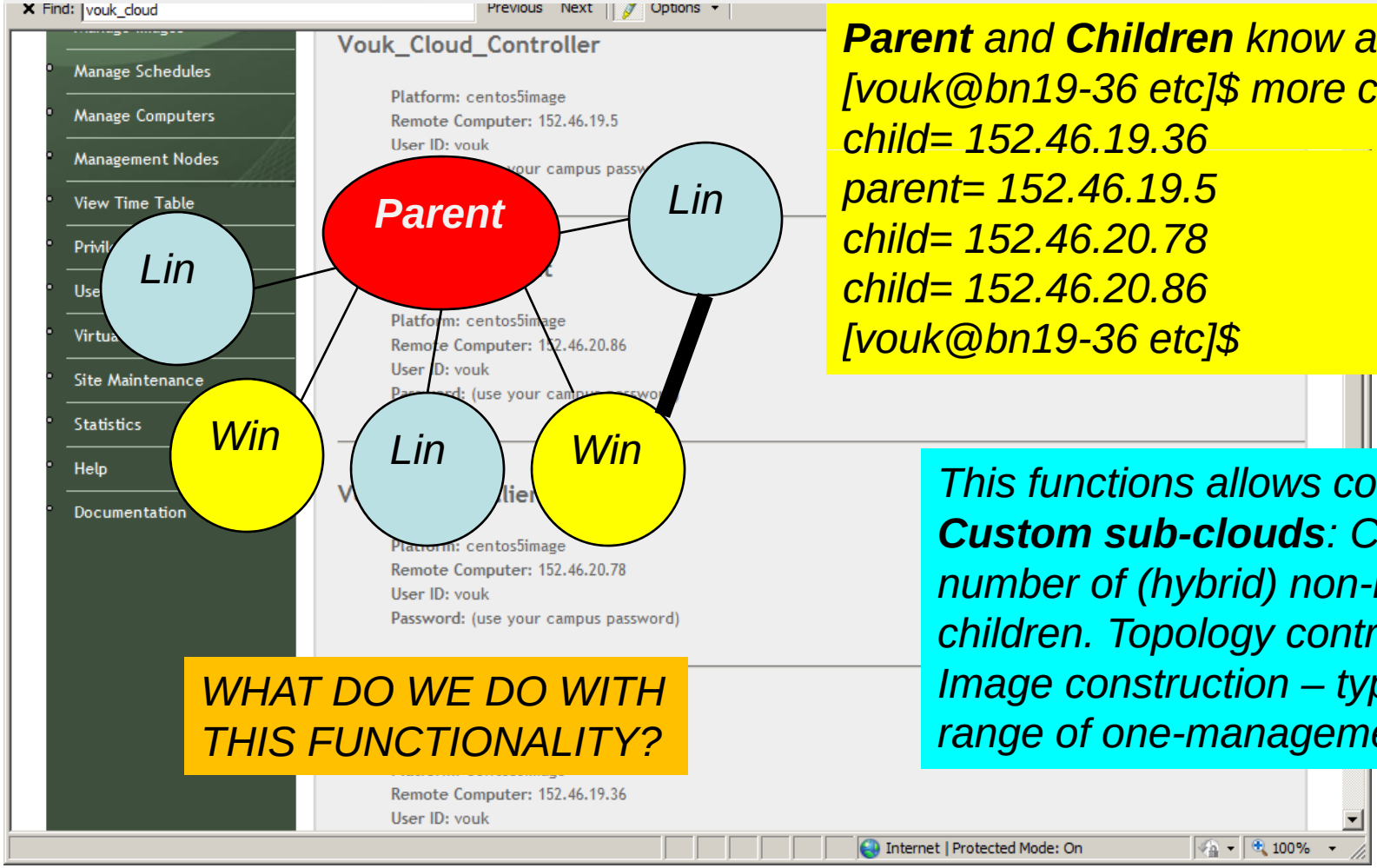
Aggregate Environments – Sub-Clouds (VCL-Cloud)

Current Reservation

You currently have the following normal reservations:

	Environment	Starting	Ending	Initially requested	Req ID	Comp ID	Management Node	IP address	Current State	Last State	Computer State
Connect! End Edit	Vouk_Cloud_Controller	Sunday, Sep 19th, 5:17 pm	Sunday, Sep 19th, 6:30 pm	Sunday, Sep 19th, 5:17 pm	1079813	1800	imgr04.dcs.mcnc.org	152.46.19.5	8	8	8

Click the **Connect!** button to get further information about connecting to the reserved system. You must click the button from a web browser running on the same computer from which you will be connecting to the remote computer; otherwise, you may be denied access to the machine.



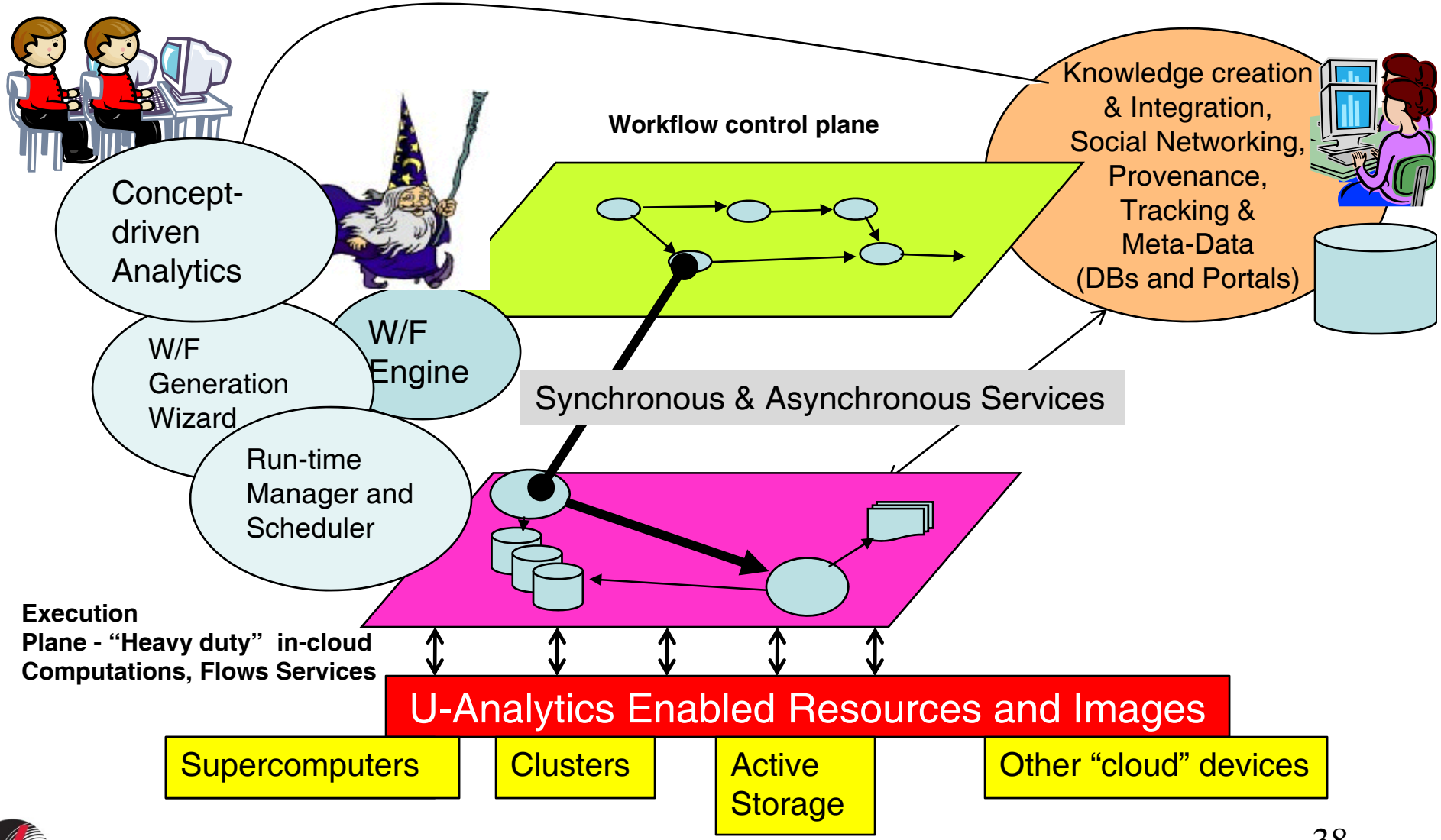
Parent *and* ***Children*** *know about each other*
[vouk@bn19-36 etc]$ more cluster_info
child= 152.46.19.36
parent= 152.46.19.5
child= 152.46.20.78
child= 152.46.20.86
[vouk@bn19-36 etc]$

WHAT DO WE DO WITH THIS FUNCTIONALITY?

This functions allows construction of ***Custom sub-clouds****: Controller + any number of (hybrid) non-recursive children. Topology control depends on Image construction – typically within range of one-management range.*

NC STATE UNIVERSITY

"U-Analytics Cloud" – under the hood



38

NC STATE UNIVERSITY

Summary

- Uncertainty analysis in many situations needs both cloud computing and workflow management support.
- VCL plans include cloud federation, analytics and uncertainty analytics sub-clouds)
- VCL experience: clouds work if utilization and reliability are high, end-user access is easy, functionality is diverse and modifiable.

http://vcl.ncsu.edu

- For a test-account on VCL, please send email

vouk@ncsu.edu