Leveraging HPCC Systems with Virtual Computing Lab

Vincent W. Freeh

Department of Computer Science

North Carolina State University

Computer Science

NC STATE UNIVERSITY

Data Intensive Curriculum

Data
- Data at scale
- Storage management
- Data warehousing
- Data format
- Encryption, compression
- Meta-data, provenance

Knowledge from information
- IR – info retrieval
- Analytics
- Inverted index
- Text processing
- Clustering and classification

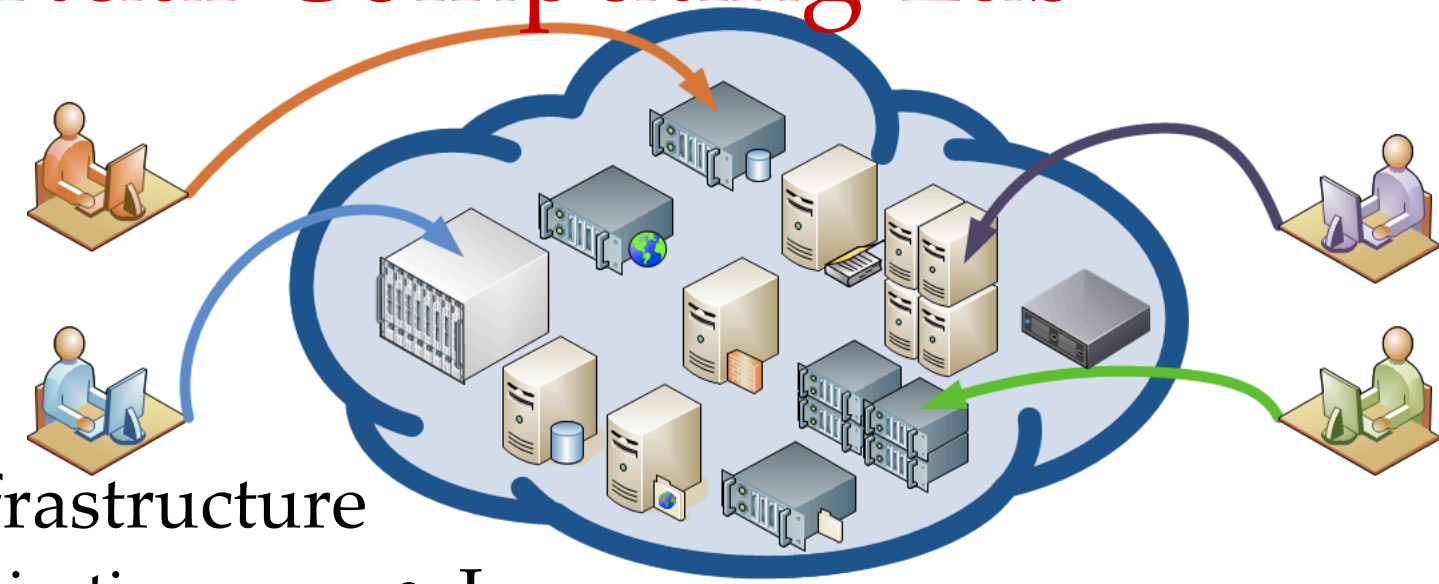
Projects

Distributed computing
- HPCC
- Hadoop
- NoSQL DBs
- Hive, Pig, zookeeper,
- BIONC/REST/AWS+

Algorithms
- MR algorithm design
- Graph algorithms

Computer Science
NC STATE UNIVERSITY

Virtual Computing Lab



- Cloud infrastructure
  - Authentication
  - Privileges
- Highly flexible
  - Time limits
  - Concurrent reservations
  - Block allocations
- Images
  - User creation
  - Bare metal or virtual machine
  - Lab machine
  - Cluster environments

Computer Science
NC STATE UNIVERSITY

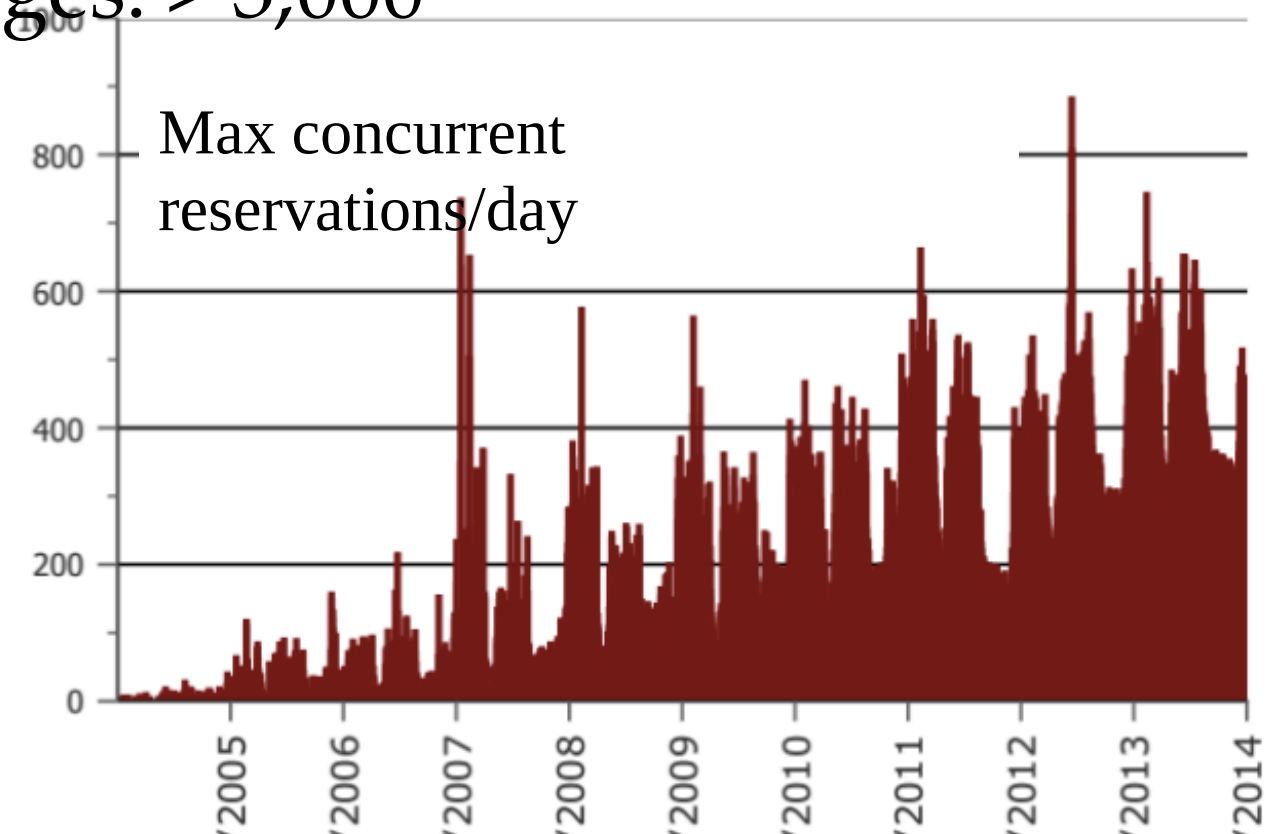
History of VCL

- Begun 2004 at NCSU
  - College of Engineering
  - Office of Information Technology
- Donated source to Apache Software Foundation 2008
  - Top-level Apache Project
- World wide
  - More than 40 installations

Computer Science
NC STATE UNIVERSITY

NCSU VCL Statistics

- Total reservations: > 1.4M
- Total hours: > 10M
- Unique images: > 3,000



HPCC on VCL

- Project: Create HPCC image on VCL
- Why
  - No setup to use HPCC
  - Experience with HPCC cluster
- Goals
  - Standalone HPCC image
  - HPCC cluster
  - Not for production (yet)

Computer Science
NC STATE UNIVERSITY

Standalone image

- To u
  - Ins
  - Do
    -
  - Cr
  - Cr
- With
  - Create reservation
  - Login

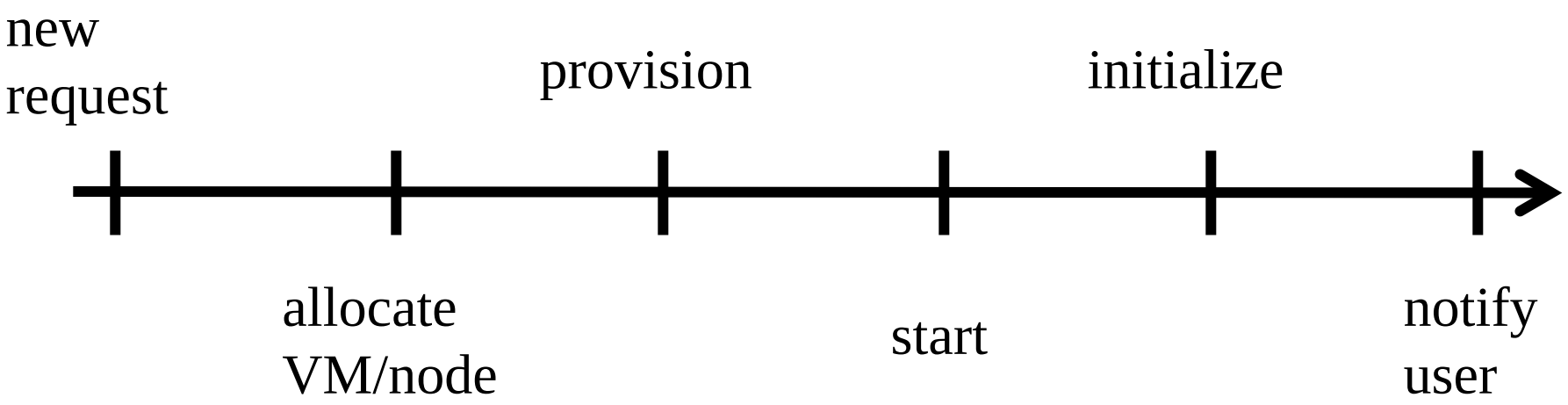
Downloads

Release	Size	Version	
HPCC VM Image 32bit Release Date: 08/18/2014 Release Notes	908.057 MB	4.2.8-1	DOWNLOAD MD5: 6be6926ebb6baec0b843f94a13b91824
HPCC VM Image 64bit Release Date: 08/18/2014 Release Notes	992.812 MB	4.2.8-1	DOWNLOAD MD5: cec84178c7f602c38afd2fb5effe2c70
HPCC VM Image 32bit Release Date: 07/28/2014 Release Notes	1016.055 MB	5.0.0-3	DOWNLOAD MD5: daf52a520eeb5ca40581728791911a6b
HPCC VM Image 64bit Release Date: 07/28/2014 Release Notes	1114.053 MB	5.0.0-3	DOWNLOAD MD5: 00ff3df03be022997ff2ae67427ad3ef

ms

Computer Science
NC STATE UNIVERSITY

VCL Timeline



Computer Science
NC STATE UNIVERSITY

NC STATE UNIVERSITY

CAMPUS DIRECTORY | LIBRARIES | MYPACK PORTAL | CAMPUS MAP | SEARCH NCSU.EDU

VCL VIRTUAL COMPUTING LAB
powered by Apache VCL

Home » Reservation System

- New Reservation
- Current Reservations
- Block Allocations
- User Preferences
- Manage Groups
- Manage Images
- Privileges
- Statistics
- Help
- Documentation

New Reservation

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

HPCC Single Node v2

Image Description:
HPCC Single Node built on RHEL 64 bit VM - v2

When would you like to use the application?

- (•) Now
- ( ) Later: Friday At 12 00 p.m. (EDT)

Duration: 1 hour

Estimated load time: < 9 minutes

Create Reservation

NC STATE UNIVERSITY

CAMPUS DIRECTORY | LIBRARIES | MYPACK PORTAL | CAMPUS MAP | SEARCH NCSU.EDU

VCL VIRTUAL COMPUTING LAB
powered by Apache VCL

Home » Reservation System

- New Reservation
- Current Reservations
- Block Allocations
- User Preferences
- Manage Groups
- Manage Images
- Privileges
- Statistics
- Help
- Documentation

Current Reservations

You currently have the following normal reservations:

			Environment
Pending... Est: 1 min remaining	Delete Reservation	More Options...	HPCC Single Node v2

This page will automatically update every 20 seconds until the *Pending...* reservation is ready.

NC STATE UNIVERSITY

NC STATE UNIVERSITY | CAMPUS DIRECTORY | LIBRARIES | MYPACK PORTAL | CAMPUS MAP | SEARCH NCSU.EDU

VCL VIRTUAL COMPUTING LAB
powered by Apache VCL

Home » Reservation System

- New Reservation
- Current Reservations
- Block Allocations
- User Preferences
- Manage Groups
- Manage Images
- Privileges
- Statistics
- Help
- Documentation

Current Reservations

You currently have the following normal reservations:

			Environment
Connect!	Delete Reservation	More Options...	HPCC Single Node v2

Click the **Connect!** button to get further information about connecting to the reserved sys
to the remote computer; otherwise, you may be denied access to the machine.

VCL VIRTUAL COMPUTING LAB
powered by Apache VCL

Cluster Reservation

This is a cluster reservation. Depending on the makeup of the cluster, you may need to use

HPCC Single Node v2

Connect to reservation using SSH (Secure Shell) on Port 22

You will need to have an X server running on your local computer and use an ssh client to
the VCL system, you will need to return to the **Current Reservations** page and click the
system. Otherwise, you may be denied access to the remote computer.

Use the following information when you are ready to connect:

Remote Computer: 152.46.20.181
User ID: vwfreeh
Password: (use your campus password)

NOTE: The given password is for *this reservation only*. You will be given a different pass
NOTE: You cannot use the Windows Remote Desktop Connection to connect to this c

NEW! You can now use SSH public key authentication to log in to SSH connections.

Issues

- Authentication
- Persistent storage

Authentication

- SSH
  - Instance is “owned” by user who created reservation
  - Can ssh into image using campus ID and password
- ECL Watch
  - Web page
  - Needs to be password protected

Authentication

- Two methods
  - LDAP
    - Not working (at this time)
    - Need to authenticate with campus LDAP server
  - .htaccess

.htaccess

- Create random password
- Create .htaccess file
- (Re)start ECL watch
- Email password to user

HPCC password

Inbox x

vwfreeh@ncsu.edu

to me

Your randomly generated password is 5gNQm3]bb
Logon to port 8010 of your machine to access ECL watch

Authentication Required

The server http://152.46.20.181:8010 requires a username and password. The server says: ESP.

User Name: vwfreeh

Password: •••••••••

Cancel | Log In

myesp – Enterprise Service

152.46.20.181:8010

HPCC Systems

EclWatch

- **Clusters**
  - Activity
  - Scheduler
- **ECL**
  - Search Workunits
  - Browse Workunits
  - ECL Playground
- **Queries**
  - Browse
  - Package Maps
- **Topology**
  - Target Clusters
  - Cluster Processes
  - System Servers
- **DFU Workunits**
  - Search
  - Browse
- **DFU Files**

Existing Activity on Servers:

ThorCluster - thor

Active workunit	State	
No active workunit		

RoxieCluster - roxie

No active workunit

HThorCluster - hthor

No active workunit

DFUserver - dfuserver_queue

No active workunit

Computer Science
NC STATE UNIVERSITY

Persistent storage

- NCSU
  - AFS storage
  - Limited
- VCL image
  - Mounts AFS as remote disk
  - Spray and despray from/to AFS
  - Done manually

Persistent storage issues

- AFS too small
- Multiple datasets
- Sharing
- Specific to NCSU

HPCC Cluster Image

- Use VCL cluster environment
  - Parent-child
  - Any number
  - /etc/cluster
- HPCC cluster configuration
  - Cluster configurations vary
  - Many parameters and options
  - Complex

Configuration

- Web page GUI
- Good for novice
- Good for persistent

Cluster configuration

- environment.xml
  - Specifies configuration
  - Easy to get wrong
  - Command line tool
- Idea
  - Create several cluster VCL images
  - Dynamically create environment.xml on each node in image
  - Start HPCC services

Computer Science
NC STATE UNIVERSITY

VCL Hooks

- Hook
  - Routine invoked by instance of image
  - Provides for dynamic configuration
  - Many hooks – at various points in the boot timeline

Example: default user

- Image is generic
- Instance has specific user and access to user's storage
- Hooks
  - Create user
  - Mount remote filesystem

Cluster Configuration

- Create environment.xml
  - Need node info for all nodes in cluster
  - Need cluster type (eg, thor-only, thor+roxie)
  - Execute command line tool
- Set up ssh keys
- Start HPCC services

Issues

- Passwordless ssh
  - Share keys during load
  - VCLs blocks general ssh
- Persistent storage
  - Even a bigger problem
- Cluster configurations
  - Create a VCL image for each configuration
  - Essentially infinitely many possible configurations
  - What are the primary clusters?

Computer Science
NC STATE UNIVERSITY

Teaching

- HPCC is a vehicle
  - Use HPCC to teach concepts
- What can be taught?
  - Applications (use ECL)
  - Distributed systems (evaluation)
  - System design (configuration)
  - Performance (identify bottlenecks)

Summary

- HPCC on VCL
  - Standalone prototype
  - Cluster prototype
- Issues
  - LDAP
  - Persistent storage
  - SSH

Computer Science
NC STATE UNIVERSITY

RESEARCH

Computer Science
NC STATE UNIVERSITY

Extending ECL with Natural Language Processing (NLP)

- GATE – open source NLP system
- Java
- Pipeline of processing resources
- Add ECL routines to create and execute pipelines

Elastic HPCC

- Elastic changes procurement (from capital to operating)
- Must effectively add or remove nodes
- Must efficiently access any data from any node