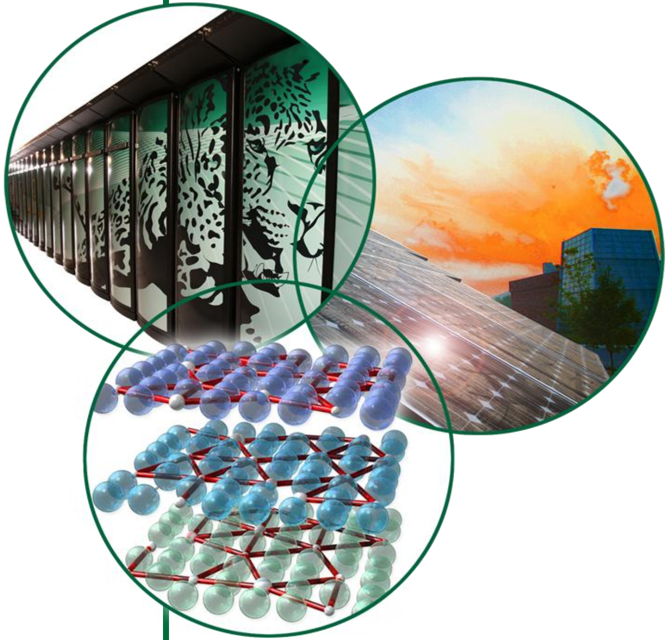
Scaling Document Clustering in the Cloud

Robert Gillen

Computer Science Research

Cloud Futures 2011



U.S. DEPARTMENT OF ENERGY

OAK RIDGE NATIONAL LABORATORY

MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

Overview

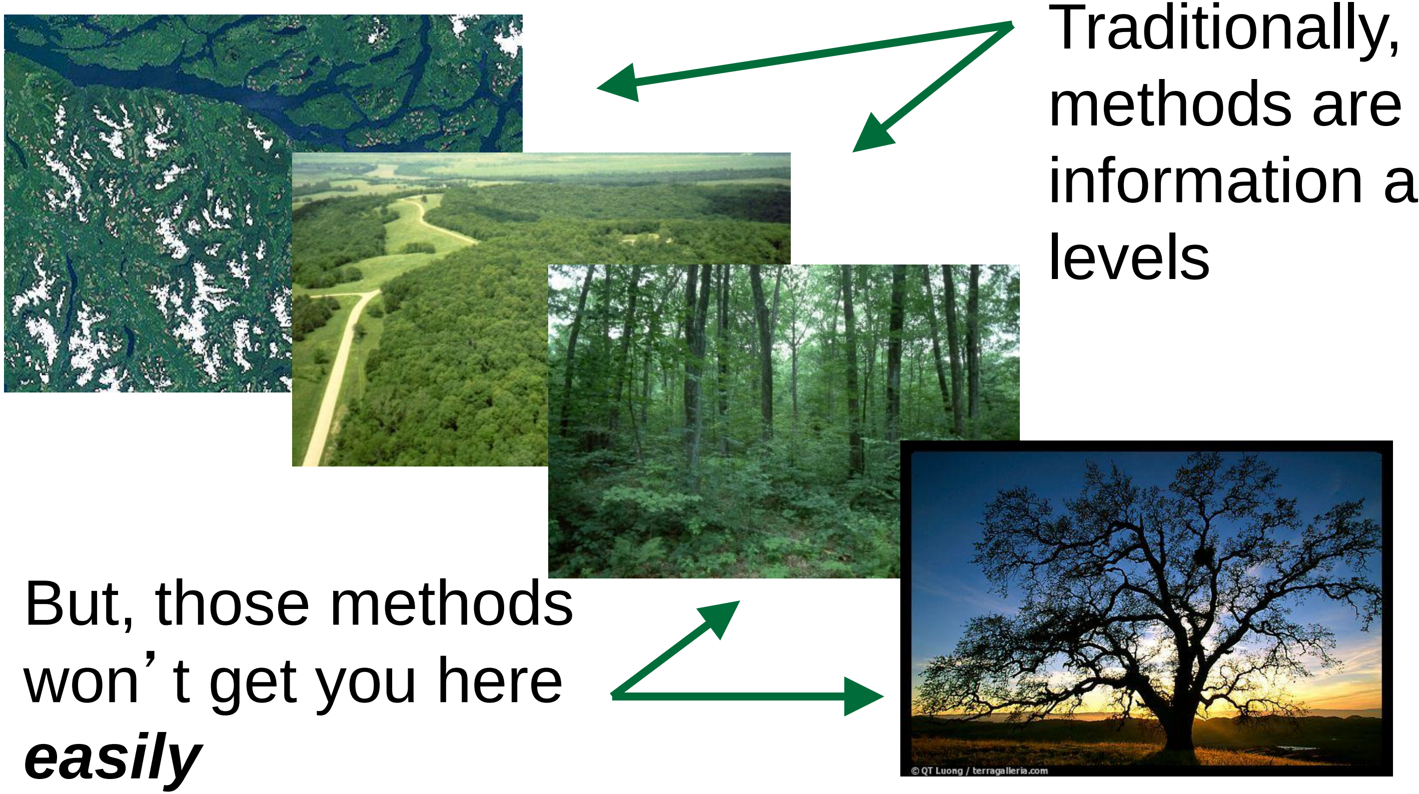
- Introduction to Piranha
- Existing Limitations
- Current Solution Tracks
- Early Results & Future Work

Challenge – What to do with mounds of data?

- What is in there?
- Are there any threats?
- What am I *missing*?
- How do I connect the “dots”?
- How do I find the *relevant* information I need?

Can't See the **Forest** for the **Trees**

Traditionally, search methods are used to find information at high volume levels



But, those methods won't get you here ***easily***

Piranha

- Ability to search *AND* analyze
  - Organize documents based on content
  - Identify similar & dissimilar documents
  - Identify duplicate and near-duplicate data
- Incorporate new data as it becomes available
- 2007 R & D 100 Award winning

Awards are based on each achievement's technical significance, uniqueness, and usefulness compared to competing projects and technologies.

Keyword Methods

Document 1

The Army needs sensor technology to help find improvised explosive devices

Document 2

ORNL has developed sensor technology for homeland defense

Document 3

Mitre has won a contract to develop homeland defense sensors for explosive devices

Term List

Army
Sensor
Technology
Help
Find
Improvise
Explosive
Device
ORNL
develop
homeland
Defense
Mitre
won
contract

Weight Terms

Term	Device
Frequency in D1	1
Frequency in D2	0
Frequency in D3	1
IDF	3/2
Term Weight D1	log(1+1)*log(3/2)
Term Weight D2	log(1+0)*log(3/2)
Term Weight D3	log(1+1)*log(3/2)

$$W_{ij} = \log_2\left(f_{ij} + 1\right) * \log_2\left(\frac{N}{n}\right)$$

Term Frequency – Inverse Document Frequency

Vector Space Model

	Doc 1	Doc 2	Doc 3
Army	1	0	0
Sensor	1	1	1
Technology	1	1	0
Help	1	0	0
Find	1	0	0
Improvise	1	0	0
Explosive	1	0	1
Device	0.28	0	0.28
ORNL	0	1	0
develop	0	1	1
homeland	0	1	1
Defense	0	1	1
Mitre	0	0	1
won	0	0	1
contract	0	0	1

An index into the document list

Textual Clustering

Vector Space Model

	Doc 1	Doc 2	Doc 3
Army	1	0	0
Sensor	1	1	1
Technology	1	1	0
Help	1	0	0
Find	1	0	0
Improvise	1	0	0
Explosive	1	0	1
Device	1	0	1
ORNL	0	1	0
develop	0	1	1
homeland	0	1	1
Defense	0	1	1
Mitre	0	0	1
won	0	0	1
contract	0	0	1

TFIDF

$$W_{ij} = \log_2\left(f_{ij} + 1\right) * \log_2\left(\frac{N}{n}\right)$$

Similarity Matrix

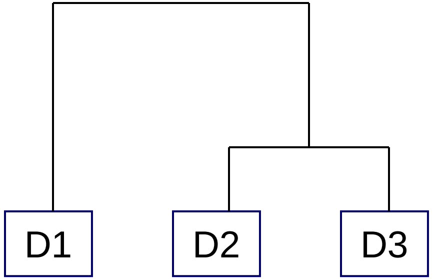
	Doc 1	Doc 2	Doc 3
Doc 1	100%	17%	21%
Doc 2		100%	36%
Doc 3			100%

Documents to Documents

Euclidean distance

$$d_2(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^{d} (x_{i,\,k} - x_{j,\,k})^2\right)^{1/2}$$

Cluster Analysis



Most similar documents

Time Complexity

$O(n^2 \mathrm{Log}\ n)$

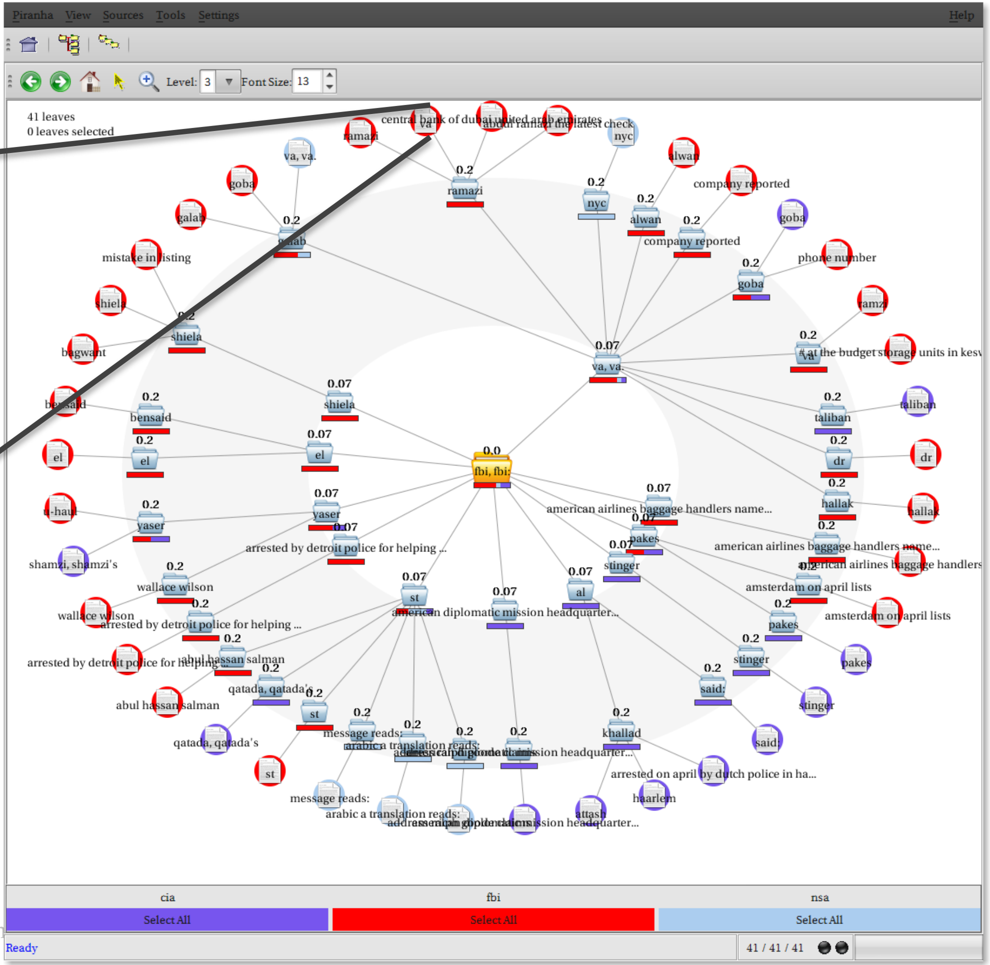
Example: Sign of the Crescent[1]

- **41 Short intelligence reports about a multi-prong terrorist attack**
- **Example:**
  - Report Date: 1 April, 2003. FBI: Abdul Ramazi is the owner of the Select Gourmet Foods shop in Springfield Mall, Springfield, VA. [Phone number 703-659-2317]. First Union National Bank lists Select Gourmet Foods as holding account number 1070173749003. Six checks totaling $35,000 have been deposited in this account in the past four months and are recorded as having been drawn on accounts at the Pyramid Bank of Cairo, Egypt and the Central Bank of Dubai, United Arab Emirates. Both of these banks have just been listed as possible conduits in money laundering schemes

[1] Intelligence Analysis Case Study by F. J. Hughes, Joint Military Intelligence College

Piranha Cluster View

Report Date: 1 April, 2003. FBI: Abdul Ramazi is the owner of the Select Gourmet Foods shop in Springfield Mall, Springfield, VA. [Phone number 703-659-2317]. First Union National Bank lists Select Gourmet Foods as holding account number 1070173749003. Six checks totaling $35,000 have been deposited in this account in the past four months and are recorded as having been drawn on accounts at the Pyramid Bank of Cairo, Egypt and the Central Bank of Dubai, United Arab Emirates. Both of these banks have just been listed as possible conduits in money laundering schemes



Existing Issues

- Memory bound
- Prior distribution approaches were troublesome
- Extant need to process larger document sets

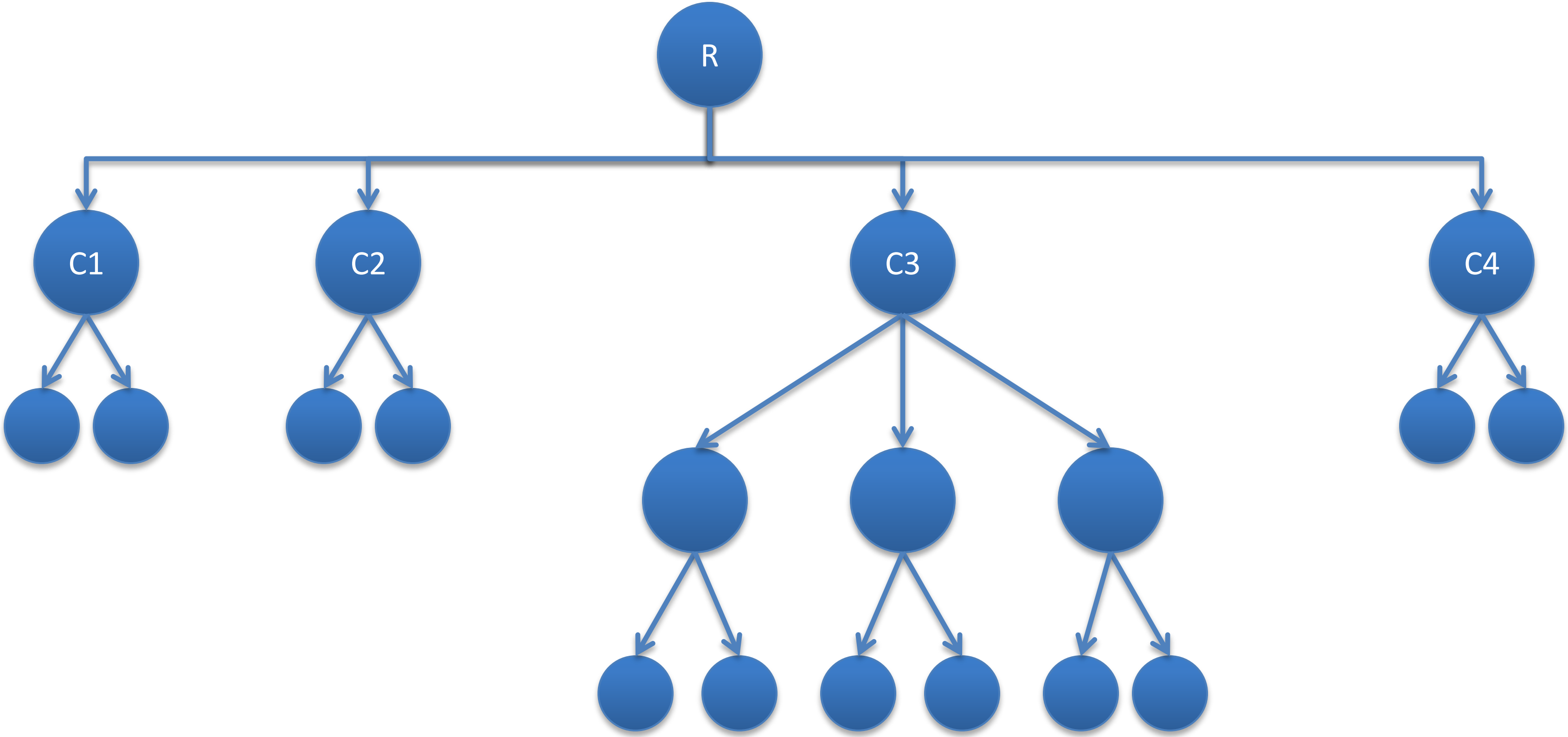
Current Solution Tracks

- Traditional HPC (Jaguar)
  - ORNL has unique capabilities in this space
- Cloud
  - New approaches may broaden the reach of the tool
    - Less-specialized hardware requirements
    - More-accessible programing/extensibility model
  - Ability to utilize core features of cloud platforms to provide key functionality

Design Tenants

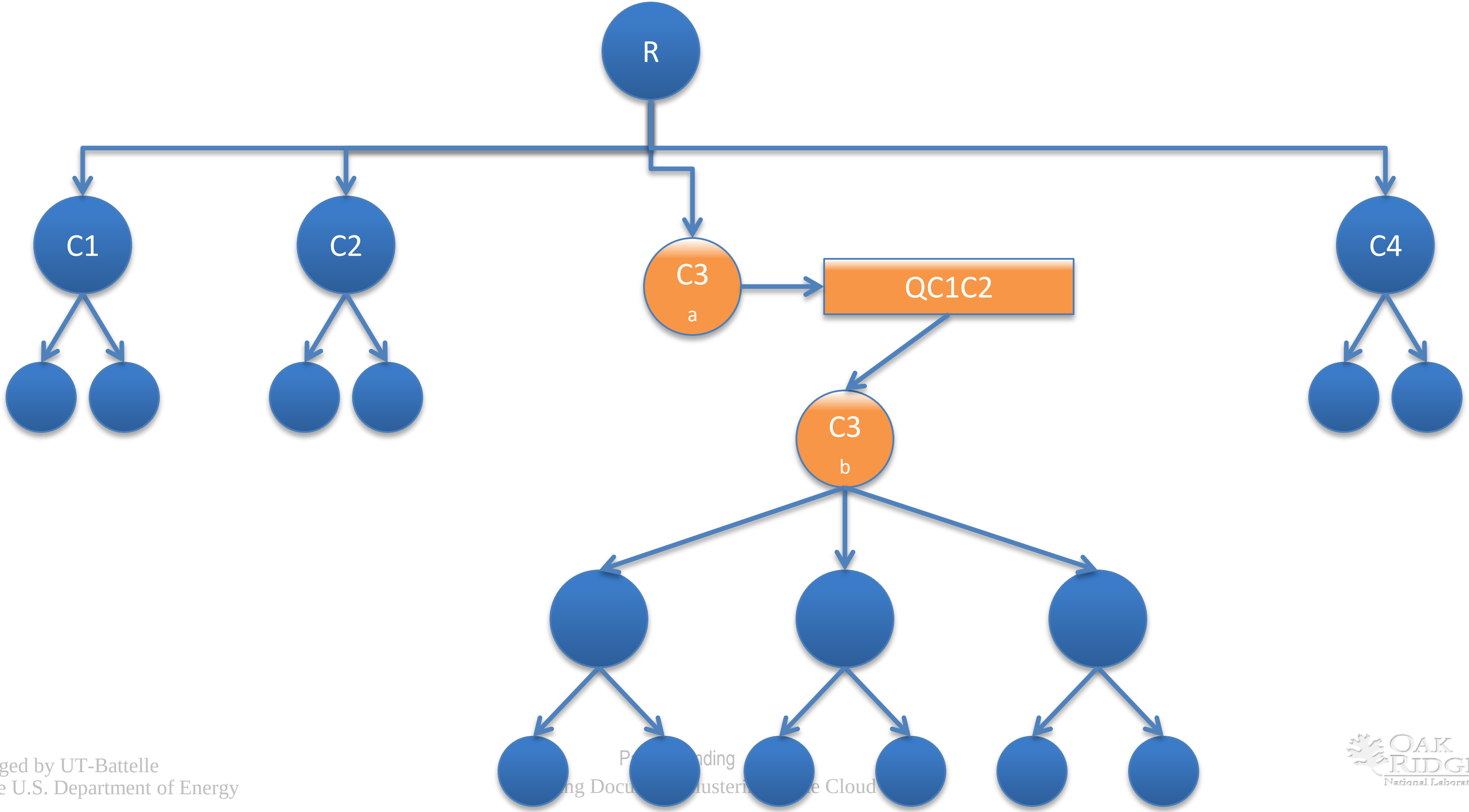
- Utilize cloud primitives wherever possible.
- Building “Environmentally Aware” algorithms… i.e. such that they are aware of the environment in which they are running.
    - Dynamically fit the platform to the problem
- Design for use in disparate environments.

Cloud Scaling Approach



Patent Pending

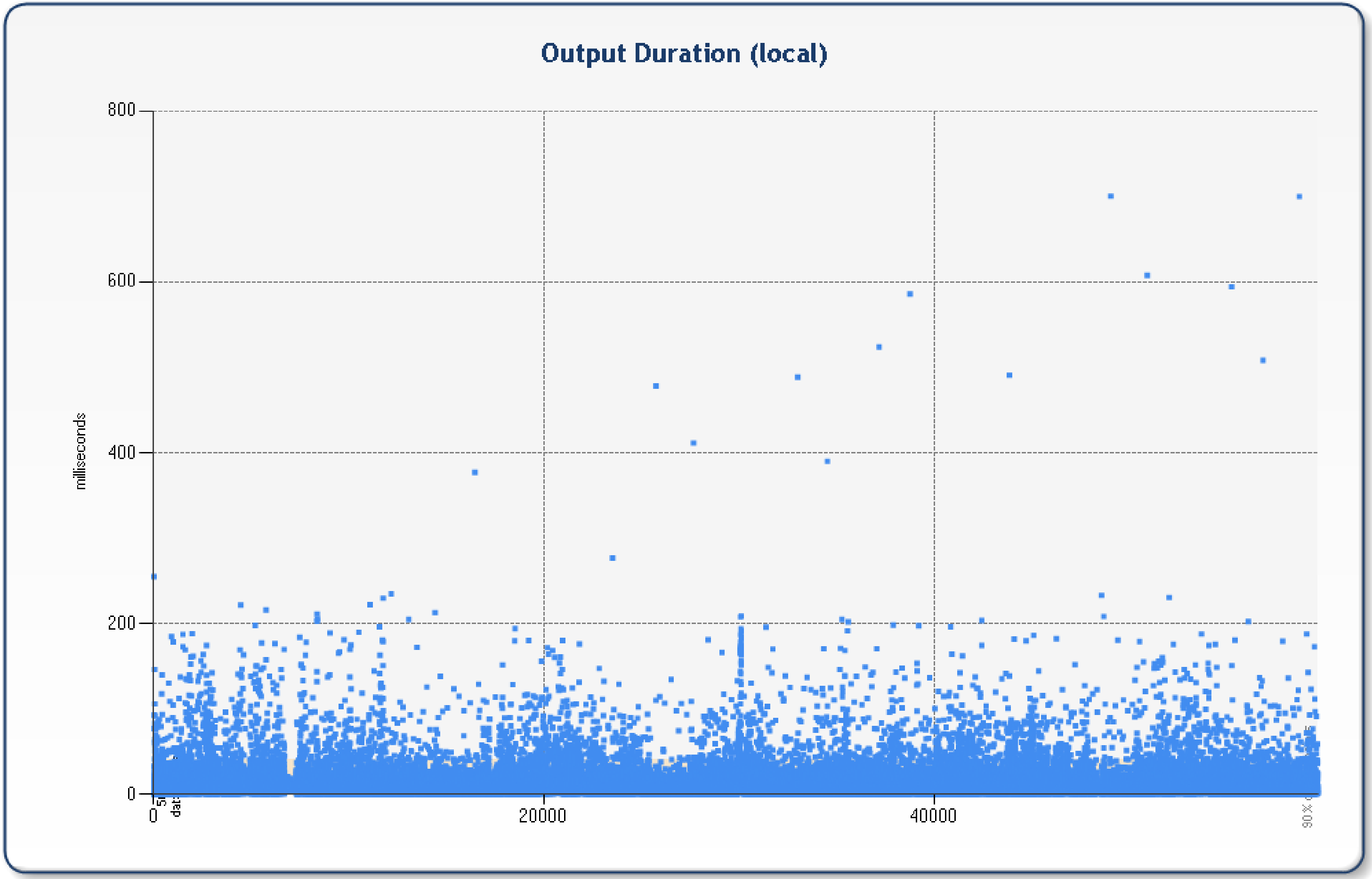
Cloud Scaling Approach



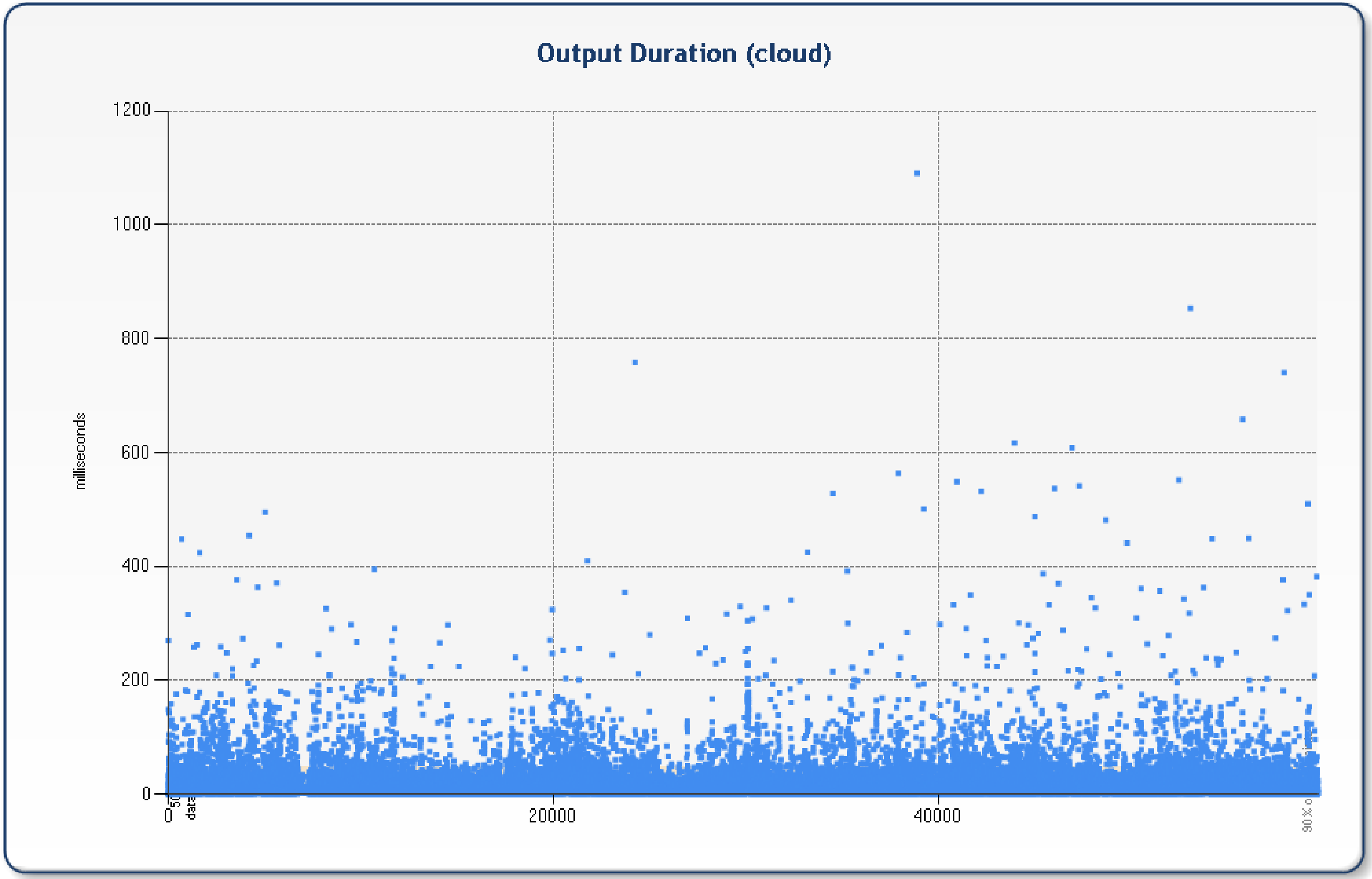
Pending Issues

- How frequently to check for memory pressure
- Work Unit Size (how many documents at a time)
- Moving from a single machine to distributed model introduces I/O delay (by definition)
- ~60K docs → increase of 2:30 – bad case, 50min/million docs

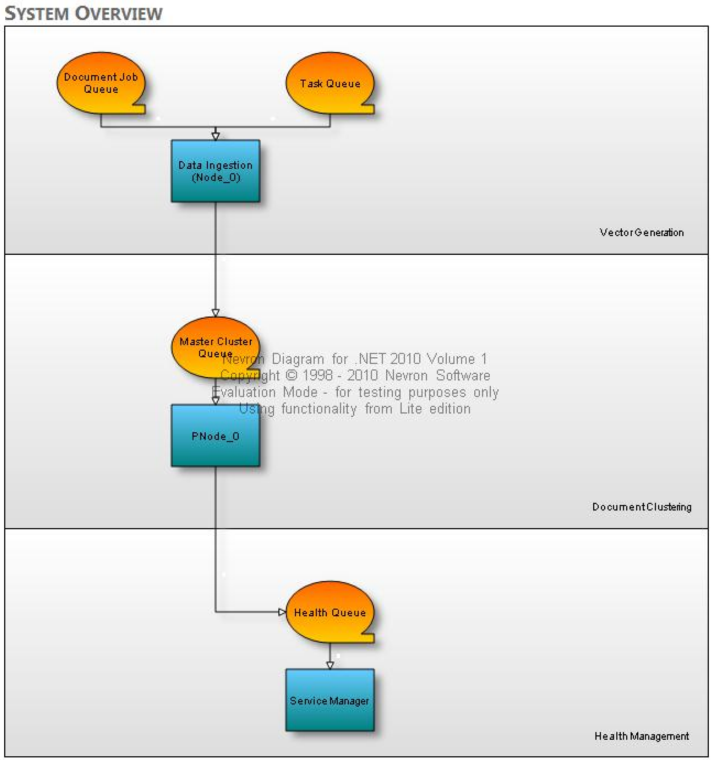
Vector Creation/Serialization (local)

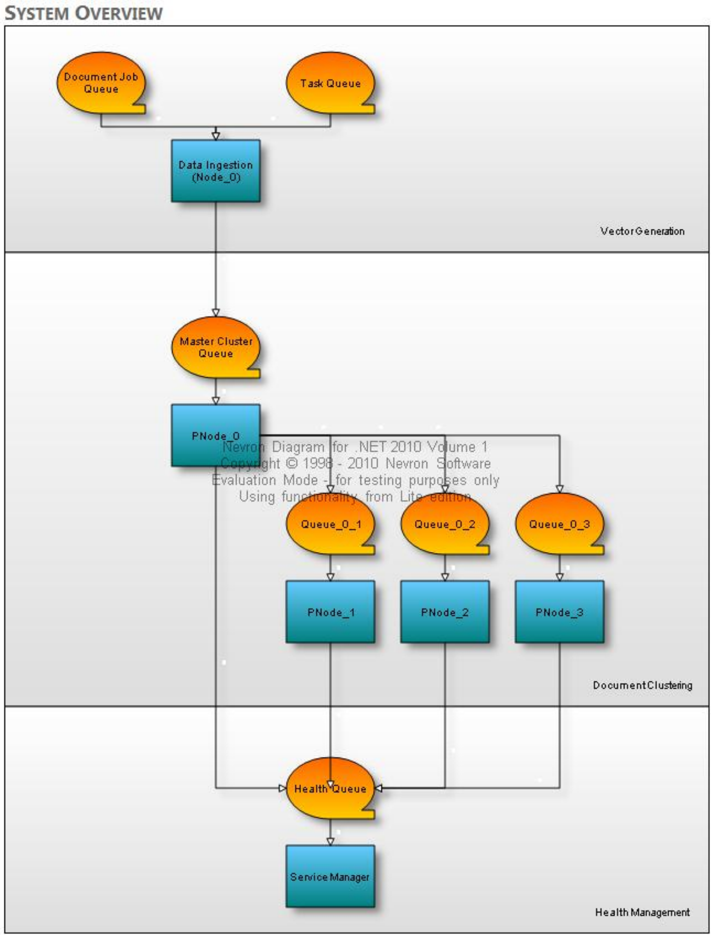


Vector Creation/Serialization (cloud)

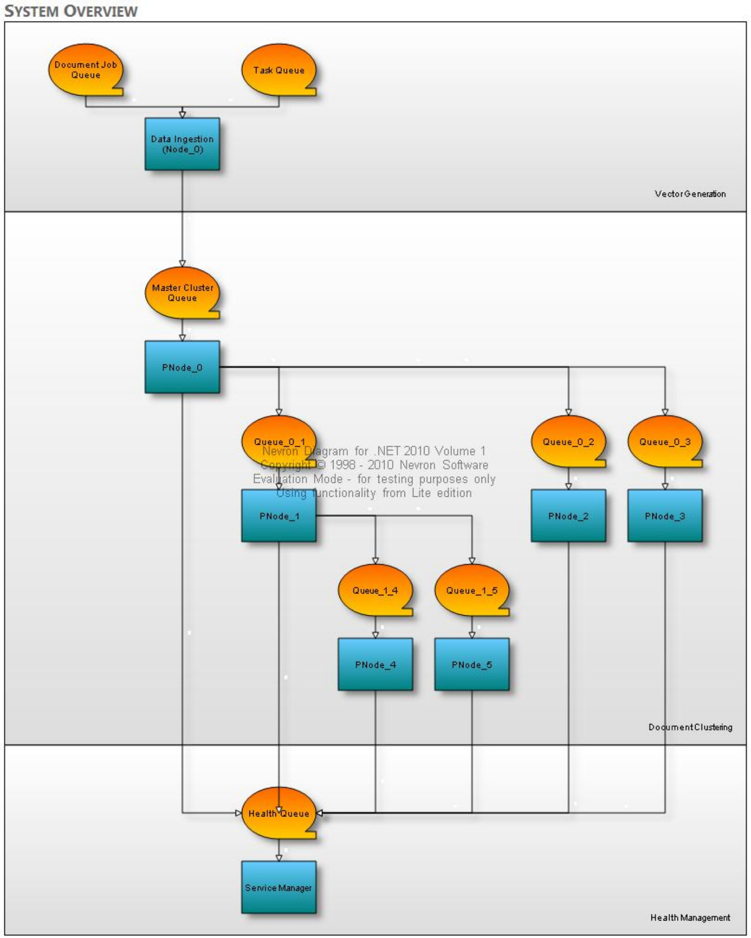


Real-Time Environment Monitoring

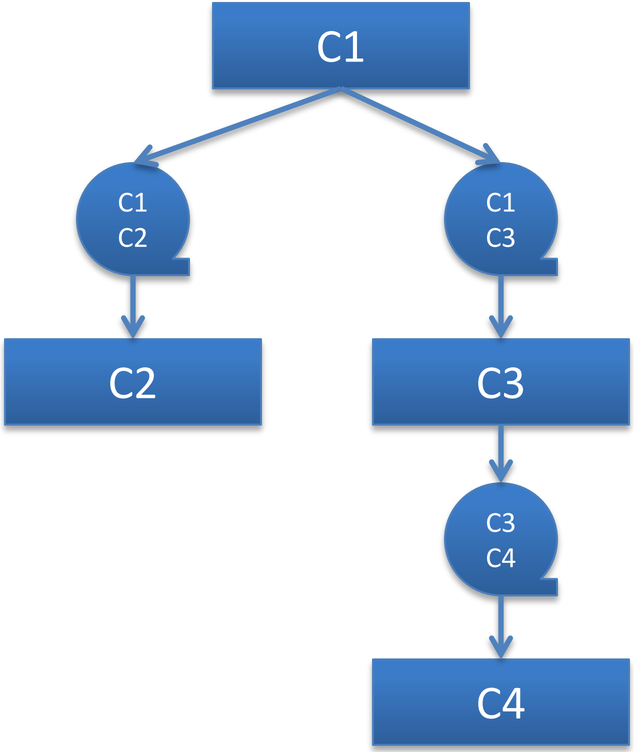




Real-Time Environment Monitoring

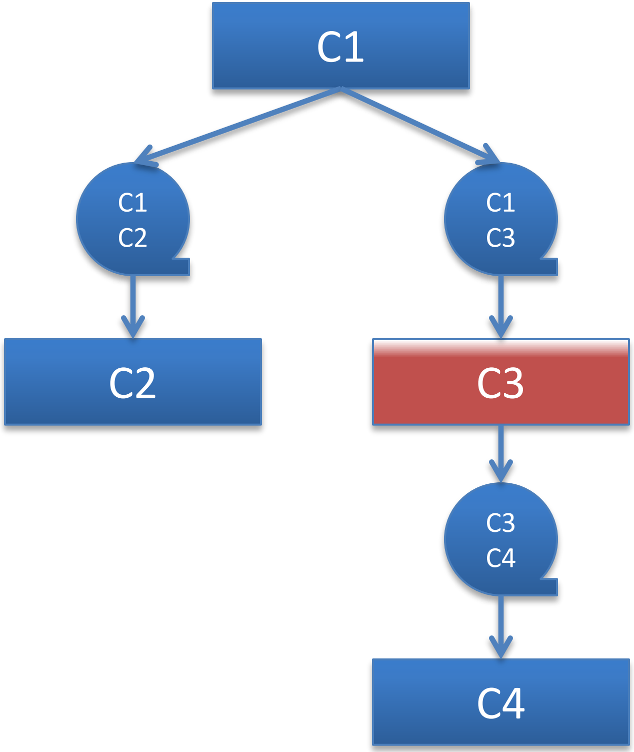


Fault Tolerance



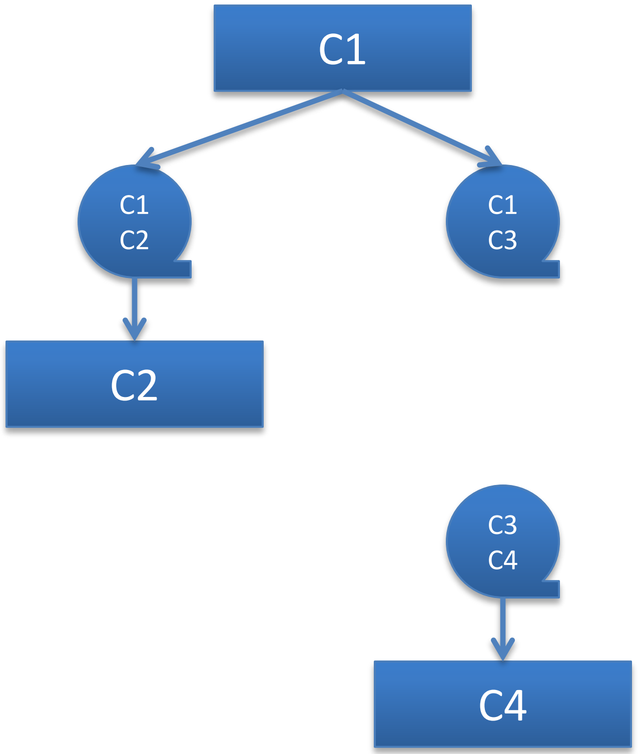
Patent Pending

Fault Tolerance

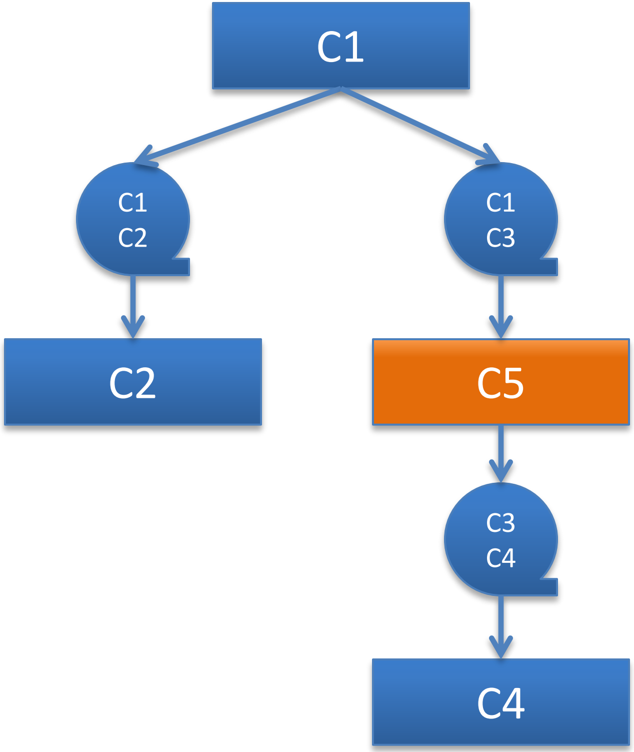


Patent Pending

Fault Tolerance



Fault Tolerance



Fault Tolerance

- Queues provide isolation for fault tolerance
- Two-phase queues are key to success
- Regular serialization of node state is key
  - Yet how often remains in question
- Not possible without programmable infrastructure provided by the cloud

Running in Different Environments

- Same core algorithm (C++ code) runs in Azure, Amazon, and on Jaguar (recompiled)
- “Scaffolding” code is cloud/jaguar specific
- Patterns used (Repository, etc) to abstract differences between various vendor storage repositories
- “Scaling” easier in Azure
- Raw control/access easier in Amazon

Early Results & Future Work

- File Packing?
- Scale vs. Stability vs. Speed
- Tuning the Work Unit Size

Questions?

Rob Gillen

gillenre@ornl.gov

@argodev