The eScience Institute:
Commoditizing Data-Intensive Scientific Analysis

Bill Howe, Phd
Senior Scientist, eScience Institute
Affiliate Assistant Professor, CSE
http://escience.washington.edu
Were you looking for: science institute

Institute of Noetic Sciences: Home Page
IONS is a nonprofit membership organization that conducts and sponsors leading-edge research into the potentials and powers of consciousness: including perceptions, beliefs ...
www.noetic.org · Cached page

Virginia Institute of Marine Science - Home
An institution that is uniquely prepared to educate the highly qualified researchers, resource managers, and educators needed for the future. Gloucester Point, Virginia.
www.vims.edu · Cached page

ICSI: International Computer Science Institute | Berkeley, California
International Computer Science Institute at the University of California, Berkeley. Research groups and technical reports.
http.icsi.berkeley.edu · Cached page

National Institute for Discovery Science (NIDS): Science institute ...
The National Institute for Discovery Science (NIDS) is a privately funded science institute engaged in research of UFOs, animal mutilations, and other related anomalous phenomena.
www.nidsci.org · Cached page
The University of Washington eScience Institute

- Rationale
  - The exponential increase in sensors is transitioning all fields of science and engineering from data-poor to data-rich
  - Techniques and technologies include
    - Sensors and sensor networks, databases, data mining, machine learning, visualization, cluster/cloud computing
  - If these techniques and technologies are not widely available and widely practiced, UW will cease to be competitive

- Mission
  - Help position the University of Washington at the forefront of research both in modern eScience techniques and technologies, and in the fields that depend upon them

- Strategy
  - Bootstrap a cadre of Research Scientists
  - Add faculty in key fields
  - Build out a “consultancy” of students and non-research staff
The University of Washington
eScience Institute

- **Funding**
  - $1M/year direct appropriation from WA State Legislature
  - $1.5M from Gordon and Betty Moore Foundation (joint with CMU)
  - Multiple proposals outstanding

- **Staffing**
  - **Dave Beck, Research Scientist:** Biosciences and software eng.
  - **Jeff Gardner, Research Scientist:** Astrophysics and HPC
  - **Bill Howe, Research Scientist:** Databases, visualization, DISC
  - **Ed Lazowska, Director**
  - **Erik Lundberg (50%), Operations Director**
  - **Mette Peters, Health Sciences Liaison**
  - **Chance Reschke, Research Engineer:** large scale computing platforms

- ...plus a senior faculty search underway
- ...plus a “consultancy” of students and professional staff
Science is reducing to a database problem

“Query the world” (Data acquisition coupled to a specific hypothesis)
“Download the world” (Data acquired en masse, in support of many hypotheses)

- Astronomy: High-resolution, high-frequency sky surveys (SDSS, LSST, PanSTARRS)
- Medicine: ubiquitous digital records, MRI, ultrasound
- Biology: lab automation, high-throughput sequencing
- Oceanography: high-resolution models, cheap sensors, satellites

“Increase data collection exponentially with FlowCam!”
The long tail is getting fatter:

- Notebooks become spreadsheets (MB),
- Spreadsheets become databases (GB),
- Databases become clusters (TB),
- Clusters become clouds (PB)

Need to think **cross-scale**
Roadmap

- Introduction
- Case Studies
  - Armbrust Lab: Marine Microbiology
  - Quinn Lab: Astrophysics Simulation
  - Lidstrom Lab: Biochemical Pathways
Case Study: Armbrust Lab
Step 1: Choose Dataset
Choose existing dataset

- Cloudburst demo reads
  - File name: 100k.br
  - Upload Date: 9/23/2009 8:20 AM

Next

Upload a new dataset

Set name: 

Browse...

Dataset Metadata
You can add arbitrary metadata to this dataset, just add your keys and values below.

[ ] key = value

+Add

Next
Step 2: Choose Reference File
You appear to have uploaded a FASTA query file, appropriate for CloudBurst. Choose a FASTA reference file below,

Choose existing dataset

- Cloudbust demo reference file
  File name: s_suis.br
  Upload Date: 9/23/2009 8:21 AM

Next

Upload a new dataset

Set name: 
Browse...

Dataset Metadata
You can add arbitrary metadata to this dataset, just add your keys and values below.

key = value  X

+Add
Next
## Step 3: Configuration

- **Allow indels in addition to mismatches**
- **Only report unambiguous best alignment**

<table>
<thead>
<tr>
<th><strong>Description</strong></th>
<th><strong>Value</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of qry and ref tuples to consider at a time in the reduce phase.</td>
<td>128</td>
</tr>
<tr>
<td>Number of mismatches / differences to allow (higher number requires more time)</td>
<td></td>
</tr>
<tr>
<td>Minimum length of the reads</td>
<td></td>
</tr>
<tr>
<td># of servers to include in the cluster (20 might be a good start)</td>
<td></td>
</tr>
<tr>
<td>Number of copies of low complexity seeds to use.</td>
<td></td>
</tr>
<tr>
<td>Suggested: # processor cores</td>
<td></td>
</tr>
<tr>
<td>Number of mappers to use. Suggested: #processor-cores * 10</td>
<td></td>
</tr>
<tr>
<td>Number of reducers to use. Suggested: #processor-cores * 2</td>
<td></td>
</tr>
<tr>
<td>Number of mappers for filtration alg. Suggested: #processor-cores</td>
<td></td>
</tr>
<tr>
<td>Number of reducers to use. Suggested: #processor-cores * 2</td>
<td></td>
</tr>
</tbody>
</table>

**Start job**
Start new Cloudburst job

**Tracked jobs**

Job ID #17:

- Map % complete: 100%
- Reduce % complete: 100%
- Results: Prepare results downloads
- Dates: 9/29/2009 10:00 PM to 9/29/2009 10:07 PM

Job ID #16:

- Map % complete: 9%
- Reduce % complete: 0%
- Dates: 9/29/2009 9:54 PM to

Job ID #15:
Waiting for cluster to start

Job ID #14:

- Map % complete: 100%
- Reduce % complete: 100%
- Results: Download raw Zip archive of results

Job ID #13:
Waiting for cluster to start
Case Study: Astrophysics Simulation
N-body Astrophysics Simulation

- 15 years in dev
- $10^9$ particles
- Months to run
- 7.5 million CPU hours
- 500 timesteps
- Big Bang to now

Source: Tom Quinn
Scalable Clustering

- Group particles into spatial clusters

YongChul Kwon, Dylan Nunlee, Jeff Gardner, Sarah Loebman, Magda Balazinska, Bill Howe
Scalable Clustering

QuickTime™ and a decompressor are needed to see this picture.

YongChul Kwon, Dylan Nunlee, Jeff Gardner, Sarah Loebman, Magda Balazinska, Bill Howe
Scalable Clustering in Dryad

YongChul Kwon, Dylan Nunlee, Jeff Gardner, Sarah Loebman, Magda Balazinska, Bill Howe
Scalable Clustering in Dryad

non-skewed

skewed

QuickTime™ and a decompressor are needed to see this picture.

YongChul Kwon, Dylan Nunlee, Jeff Gardner, Sarah Loebman, Magda Balazinska, Bill Howe
Case Study: Biochemical Pathways

Lidstrom Lab
Ex: Biochemical Pathway Visualization

Serine Cycle

source: Betsy Skovran
Case Study: Computational Oceanography

NSF Ocean Observatories Initiative
Unstructured Grids

Columbia River Estuary

“unstructured grid” model complex domains at multiple scales simultaneously

....but complicate processing

red = high salinity (~34psu)
blue = fresh water (~0 psu)
GridFields: Query Algebra

Algebraic Manipulation of Scientific Datasets,
B. Howe, D. Maier, VLDBJ 2005

H₀ : (x,y,b)  V₀ : (σ)

A

B

C

apply(0, z=(surf – b) * σ)

bind(0, surf)

restrict(0, z > b)

color is salinity
Conclusions

- So many great projects!
  - Always looking for partners -- we inoculate ourselves against “NIH” syndrome
  - oceanography, astronomy, microbiology, civil engineering, health sciences, many more
- No one-size-fits-all solution
- Cross-scale is important
- # of apps scale linearly with # of bytes
  - So we need to handle that too
- Visualization Database Management Systems are an important open problem

Ask me about this…. 