dynamicBLAST on SURAgrid:
Overview, Update, and Demo

John-Paul Robinson
Enis Afgan and Purushotham Bangalore
University of Alabama at Birmingham

SURAgrid All-Hands Meeting
March 15, 2007
Washington, DC

What is BLAST?

- BLAST stands for Basic Local Alignment Tool
- BLAST is search tool to find gene sequences in gene databases
- The search is really a statistical analysis of the “closeness” of a gene sequence, not a simple string pattern match
- There are many implementations
- NCBI (National Center for Biological Information) has much more information

How Do I Run BLAST?

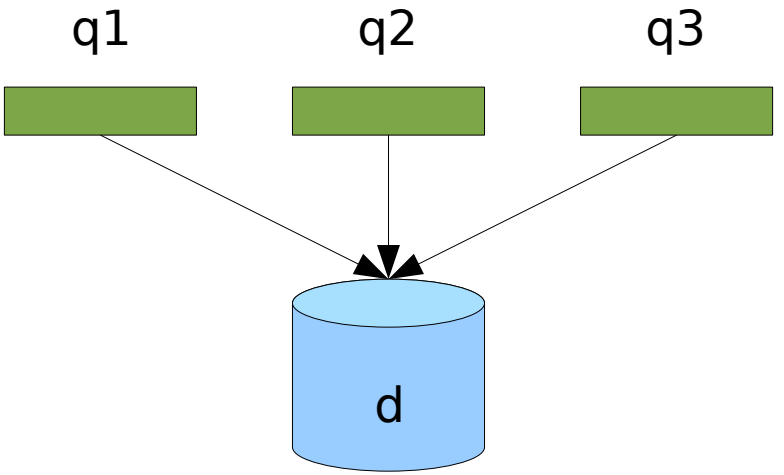
- Find a gene sequence that interests you (research)
- Select a BLAST program that implements the search type you're interested in (nucleotide, protein)
- Select a sequence database for the organism of interest (yeast, E.coli, ...)
- Begin the search
- Analyze the statistical results
- Repeat

BLAST Requirements

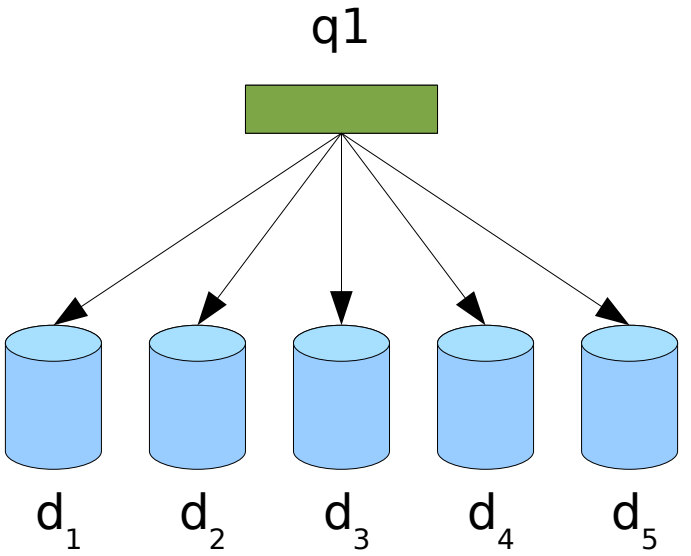
- BLAST application must be installed
  - Many implementations
  - NCBI BLAST packages for many common platforms
- Install target organism databases
  - NCBI publishes many
  - Size is 100's MB and growing
- Match your computer configuration to the application demands

How to Run BLAST Faster?

- Query Splitting
  - Many queries as fast as possible
  - Database reads affect speed
  - Low IPC makes COTS clusters ideal



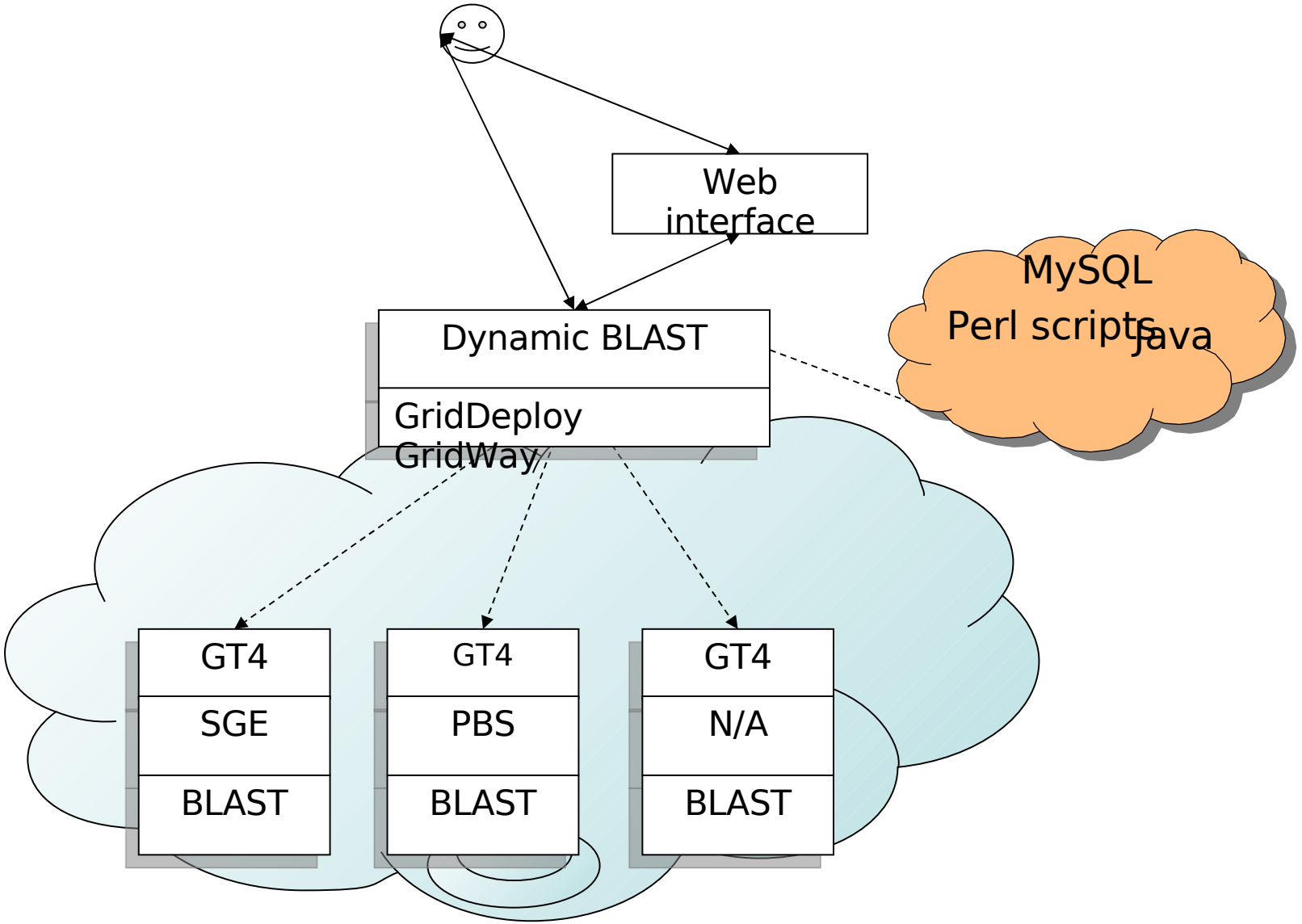
- Database Splitting
  - One query as fast as possible
  - Database chunks improve reads
  - Much IPC for boundary condition on statistical analysis (MPI)
  - SMP is best



Can Grids make BLAST Faster?

- Focus on Query Splitting because of the low inter-process communication (IPC)
- Focus on WorkFlow management to maximize query throughput and resource utilization
- dynamicBLAST takes these approaches

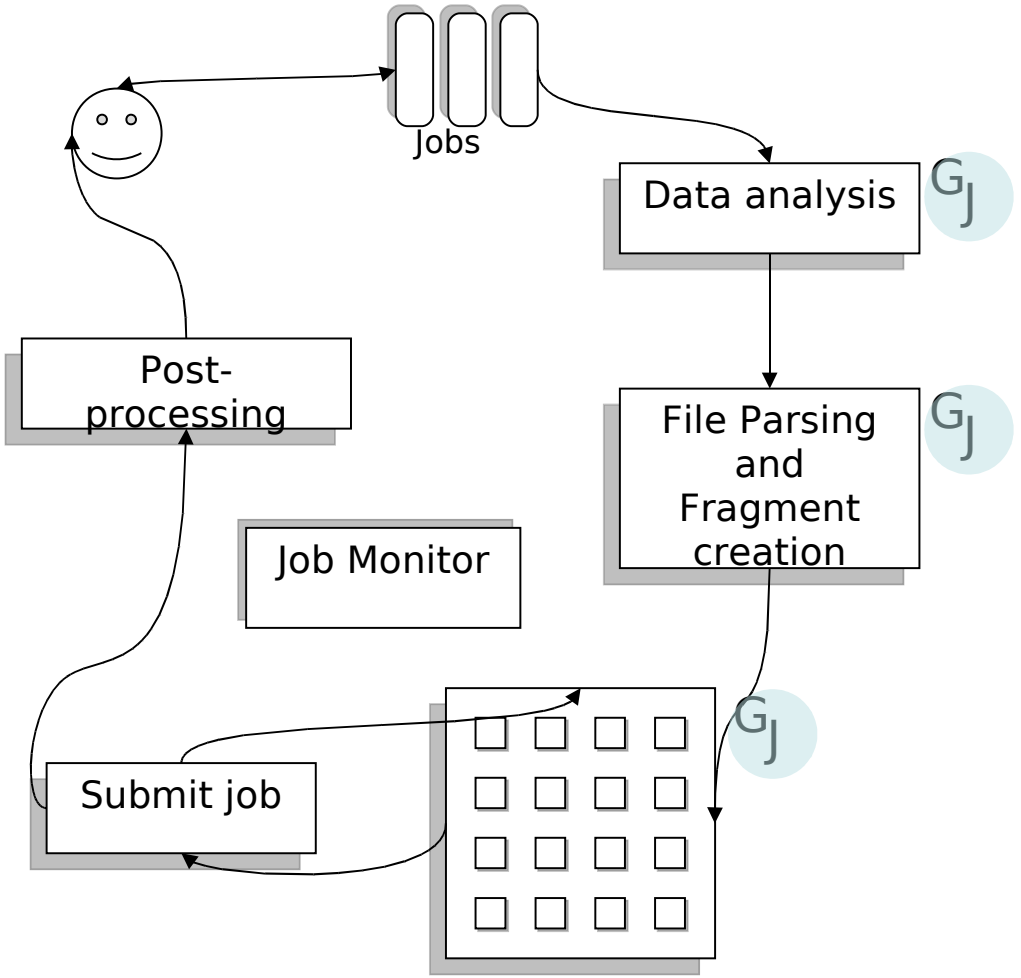
dynamicBLAST architecture



dynamicBLAST Operation

- Accepts user query list
- Maintains resource & application configuration in local database
- Processes queries in a 3-stage workflow
- Uses GridWay to manage workflow and job distribution across grid nodes

dynamicBLAST workflow



Stage 1: Analysis

- Analyzes job processes user query list
- Helps with algorithm selection
- Determines granularity of splitting
- Uses GridDeploy to manage the analysis

Stage 2: Fragmentation

- Fragmentation job divides user query list
- Potential to divided query list based on resource capabilities
- Currently just divides query list across available resources

Stage 3: BLAST Searchs

- Master / Worker model
- Assigns queries to available resources
- All reporting resources used
- Loops through all queries until done
- Takes resources as they become available

Deployment on SURAgrid

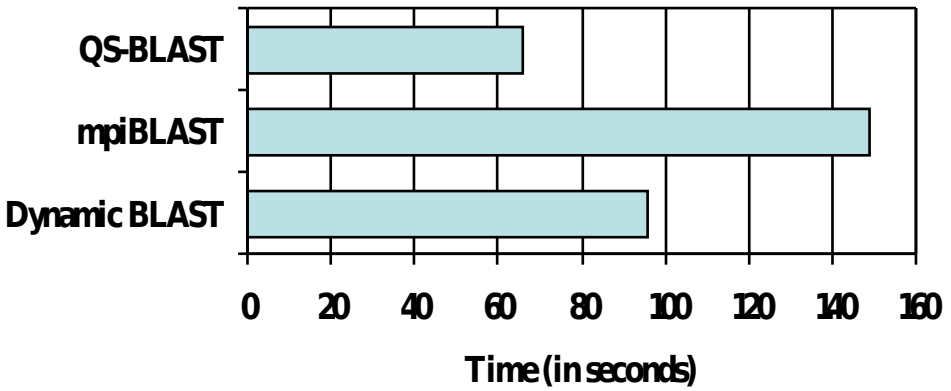
- Incorporate ODU resource Mileva: 4-node cluster dedicated for SURAgrid purposes
- Local UAB BLAST web interface
  - Leverage existing UAB Identity Infrastructure
  - Cross certification with SURA Bridge CA
- Authorization on a per-user basis
- Initial Goals:
  - Solidify the execution requirements of DynamicBLAST
  - Perform scalability tests
  - Engage researchers further in the promise of grid computing

Demo

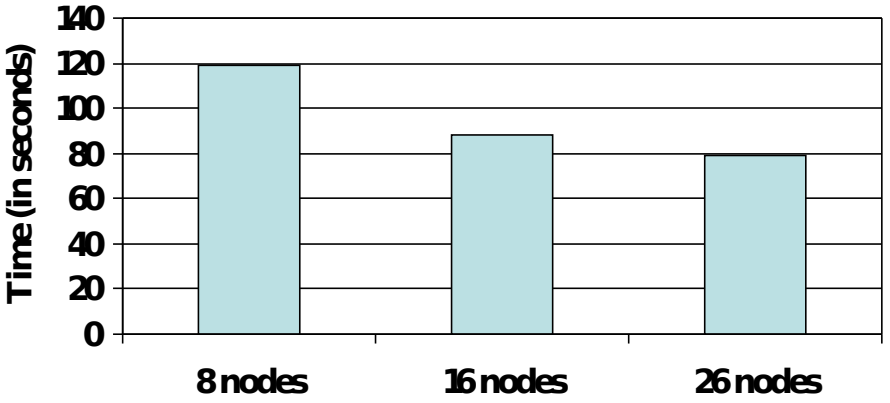
[afgane@everest00 drmaa]$ ./dblast51.sh
[2007/03/15 08:27:08] Started Dynamic BLAST.
[2007/03/15 08:27:08] Using job properties file: props.properties
...
[2007/03/15 08:27:08] Analisys job successfully submitted, job ID: 64
[2007/03/15 08:28:01] Job 64 finished regularly with exit status 0
...
[2007/03/15 08:28:01] Total number of fragments to be created: 10
[2007/03/15 08:28:01] Fragmentation job successfully submitted, ID: 65
[2007/03/15 08:29:03] Job 65 finished regularly with exit status 0
...
[2007/03/15 08:29:03] Gathering resource information...
[2007/03/15 08:29:29] Succesfully accesed current resource data.
availableResourceList=[[mileva.hpc.odu.edu, 4], [titanic.hpcl.cis.uab.edu, 4],
[olympus.cis.uab.edu, 2], [everest.cis.uab.edu, 18]]
...
[2007/03/15 08:29:29] Remaining num jobs (frags) to submit: 10
[2007/03/15 08:29:29] Creating job(s) for mileva.hpc.odu.edu with 4 available node(s).
[2007/03/15 08:29:29] Creating job(s) for titanic.hpcl.cis.uab.edu with 4 available node(s).
[2007/03/15 08:29:29] Creating job(s) for olympus.cis.uab.edu with 2 available node(s).
 exit_time=08:30:37
Job usage (for jobs [66, 67, 68, 69]) on titanic.hpcl.cis.uab.edu:
 ...
****
All fragments have been processed.
****
...

dynamicBLAST performance

- Comparison of execution time between query splitting BLAST,
  Dynamic BLAST and mpiBLAST
  for searching 1,000 queries against yeast.nt database



- Comparison of execution time of Dynamic BLAST on variable number of nodes for 1,000 queries against yeast.nt database



dynamicBLAST Experiences

- Address how to split workflow across grid nodes
    - Many jobs have distributable sub-tasks
    - Manages coordination of dependencies
- BLAST tickles many problems of grid space
    - Data distribution
    - Coordinated Execution
    - Holy Grail – MPI on grid :)

SURAgrid Experience

- Combine Local needs with regional needs
- No Need to Build Campus Grid in Isolation
- Synergistic Collaborations
- Leads to Solid Architectural Foundations
- Leads to Solid Documentation
- Extends Practical Experience for UAB Students

Next Steps

- Integrate dynamicBLAST with familiar BLAST web interface
- Integrate web interface with UABgrid 2.0 collaboration environment
  - Leverage Shibboleth, GridShib, and myVocs to federate web inteface
- Extend the number of resources running BLAST
  - ODU integration helps document steps and is model for additional sites
- Explore other applications relevant to UAB research community

References and Contacts

- Purushotham Bangalore, Computer and Information Sciences, UAB puri@uab.edu
- Enis Afgan, Computer and Information Sciences, UAB afgane@uab.edu
- John-Paul Robinson, HPC Services, Office of the VP of IT, UAB jpr@uab.edu

- http://uabgrid.uab.edu/dynamicblast