All-Pairs: An Abstraction for Data-Intensive Computing

Christopher Moretti, Jared Bulosan, Douglas Thain, and Patrick J. Flynn

To attack very large scientific problems, users must branch out beyond computation on a single machine. Most scientists are not grid experts. Thus, they make poor decisions when faced with grid software options, job infrastructure, data transfer, and other parameters. The users get poor performance, and the grid infrastructure gets abused.

We propose that abstractions can provide better performance and make grids easier to use for data intensive workloads. An example of an abstraction is All-Pairs.

Why an Abstraction?

- Hard problems can be defined in terms of a simple model.
- Abstractions hide messy grid details from the user.
- Thus, preventing unintended consequences of their poor choices.
- The system can optimize the structure of the workload, the tools .

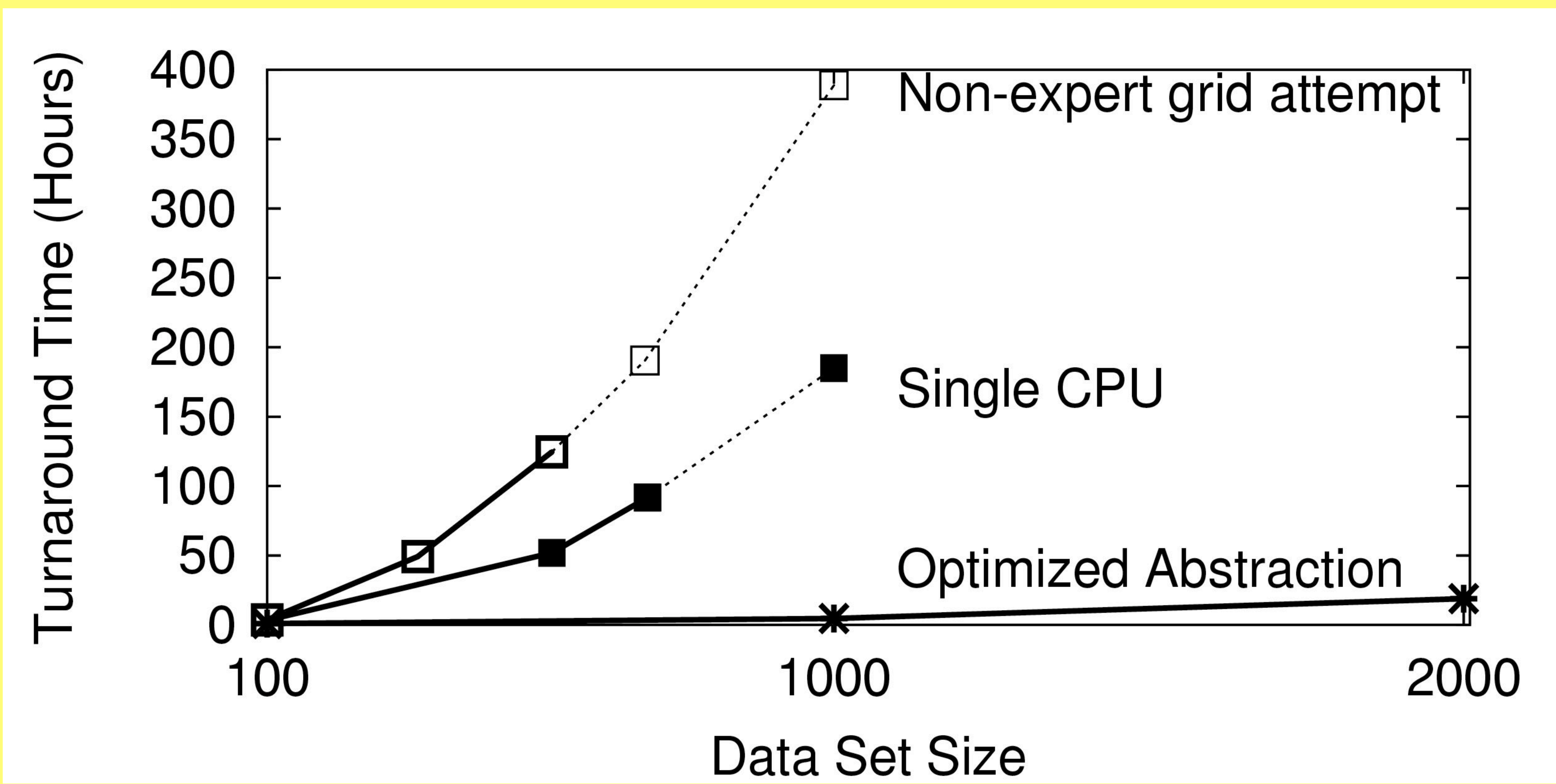
Why is All-Pairs Hard?

- What challenge does a bigger problem size pose?
  - Big problems mean higher minimum requirements, so fewer usable resources.
- How many CPUs to use?
  - Using more CPUs requires more data transfer in exchange for more parallelism.
- How much work do we give each machine at a time?
  - Large job sizes amortize grid overhead, but mean more evictions and badput.
- What else is different than the same problem done locally?
  - Common applications weren't designed for the grid, but users still want them.

What Comes Next?

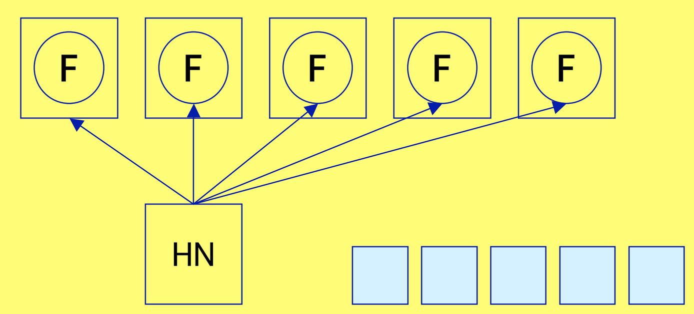
- Active Storage for data sets too large for any single node.
- A suite of abstractions and a language to define workloads.

Initial Results on Biometric Workload

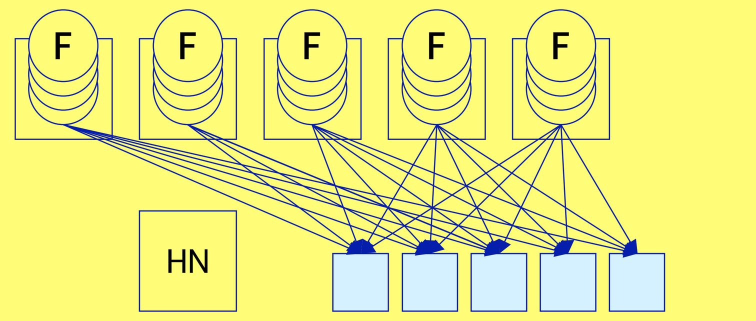


Non-Expert Grid User with 500 CPUs

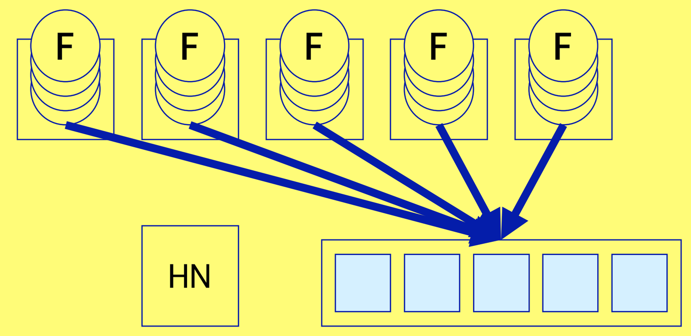
Try 1: Each F is a batch job.
Failure: Dispatch latency >> F runtime.



Try 2: Each row is a batch job.
Failure: Too many small ops on FS.

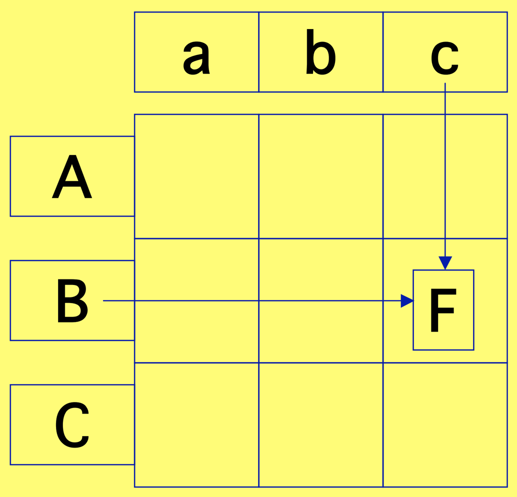


Try 3: Bundle all files into one package.
Failure: Everyone loads 1GB at once.



Try 4: User gives up and attempts to solve an easier or smaller problem.

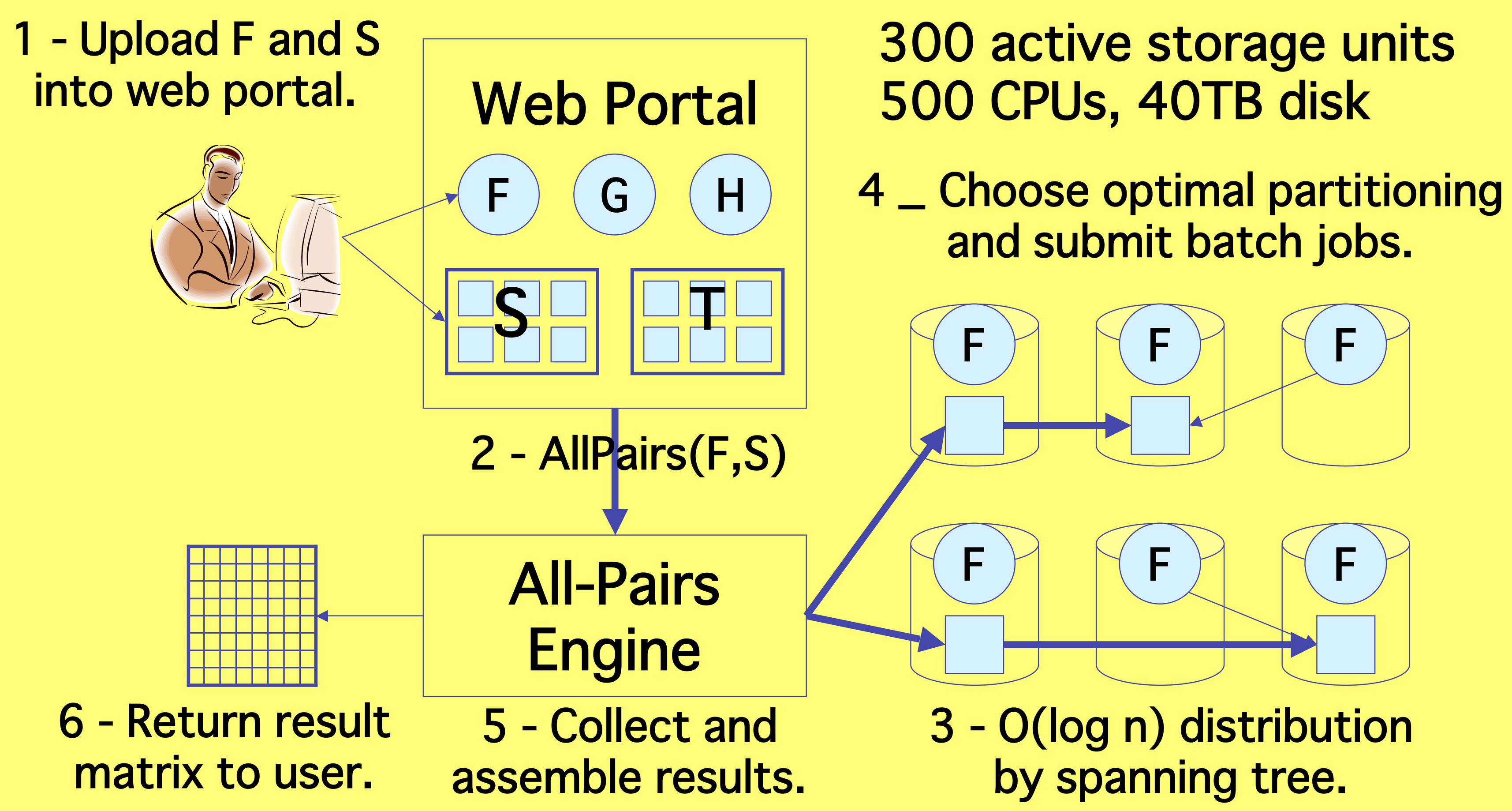
All-Pairs Abstraction



General All-Pairs Problem
A function applied to all possible pairs of elements from two sets.

Biometric All-Pairs Face Comparison
Other applications include data mining and computational biology.

Notre Dame All-Pairs Production System



Presented at the IEEE/ACM International Conference on Grid Computing, 20 September 2007.
This work was supported by NSF Grant CCF-0621434 "Deconstructing Clusters for High End Biometric Applications".
Web address: http://www.cse.nd.edu/~ccl