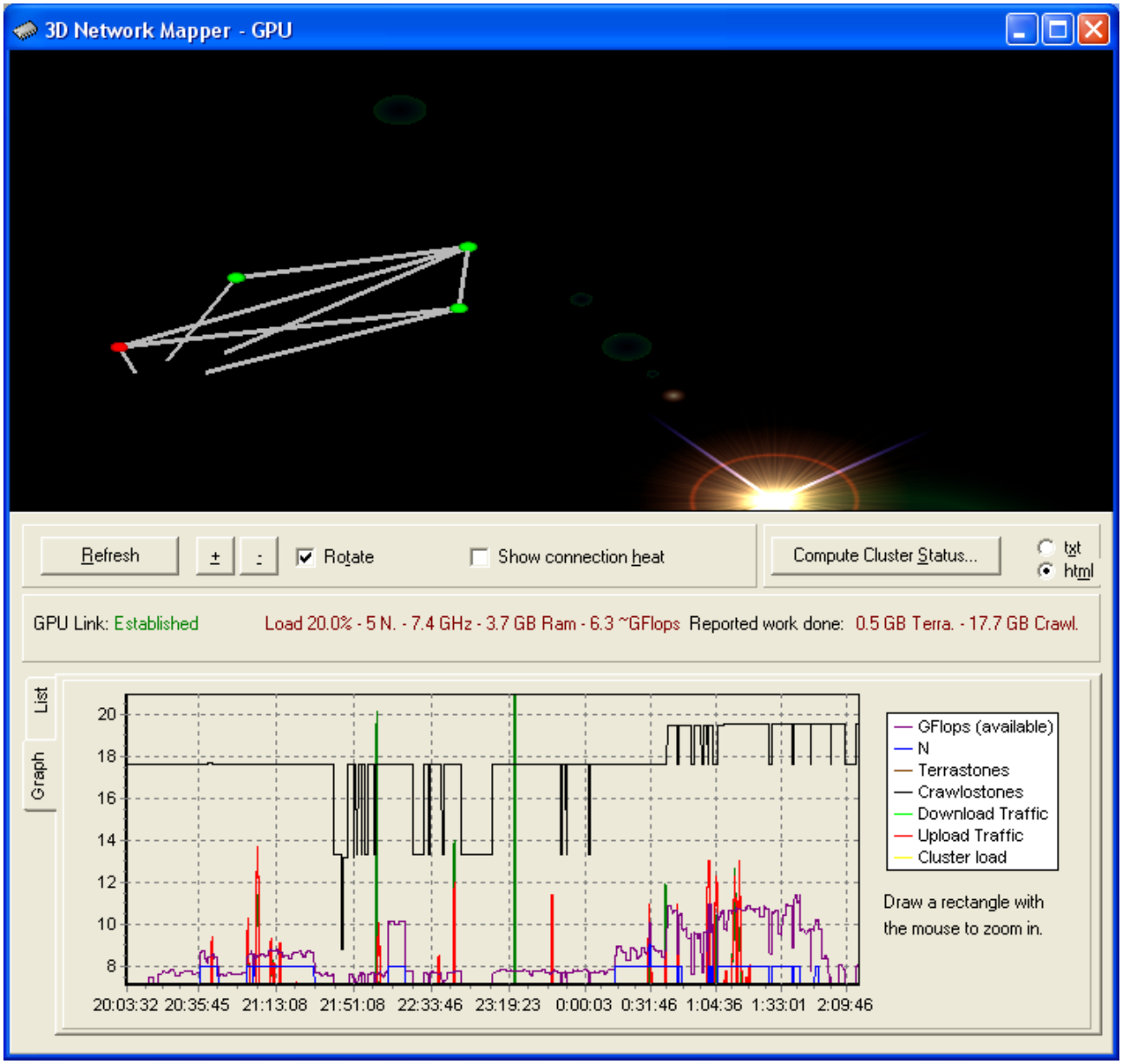
News on the GPU Search Engine (Distributed Web Indexing)

by René Tegel – september 25. 2005

The first beta run of the renewed GPU Search Engine went smoothly. Changes to the database layer made the crawler much more efficient. Users crawled hundreds of megabytes easily, internet connection being the primarily limiting factor.

I want to express great thanks to lwm ond paul for running their boxes over a long time. Together with two of my boxes we built a big database (over 20 gigabytes) in a timespan of about 2-3 weeks. The amount of internet traffic is much larger, there are no statistics at the moment, but it might be at least a factor 10 or 20 larger. 250GB of transported http traffic and indexed html files is not an underestimation. Only a fraction of the html is actual textual data, generally somewhere between 5 and 10%. Unofficial numbers say the web as 'plain text' data is at least about 150,000GB large, so we have a long way to go ;)



The sqlite database format is consired ok for the moment, no real big changes are expected. The database of the mysql version is extended with a reverse index, this is under construction, but altering tables is easily done with mysql (in contrast to sqlite).

So, we arrive at the next stage of the project. The crawlers have indexed a not negligible part of the web, having hits for most common web pages. We already set up a reversed index linked to a web page. On new queries from a www visitor on http://search.dubaron.com, the crawlers actively respond and start sending results back. This works well, but the gpu network appears to be a limiting factor. The amount of traffic explodes and causes problems for nodes with reduced bandwidth when 4-5 nodes start sending back data on a popular or common search term.

Anothor issue is the crawler itself. I encountered a few domains that were able to mislead the crawlers and guided them to a pitfall, by creating an almost endless number of subdomains for any odd keywords you could think of (really. complete dictionairies.) The fix is easy: blackklist them. If you are running a crawler now or in the future and notice such behaviour or other undesired behaviour, please report.

A comparable but less harming issue is the amount of urls that certain industries, especially porn industry, is able to create. it is not a real problem and frontends may filter it out, but.. i'd like to see the crawler being more interested in scientific material, like indexing sites as nasa and wikipedia properly, than returning a billion hits on the keyword sex.

What is on the todo and wish list:

- The crawler must have some more configuration: blacklist, whitelist etc, maybe other stuff, that allows more control on the crawler instead of modifying the binary.
- I'd like to introduce favorites: a list containing subjects / keywords which you would like to specialize your crawler in. If it will not go to NASAby itself, you can make it do so. Other niches dito. Ideas about this subject are welcome.
- I'd like to add a pdf index functionality. pdf-to-text libraries are publicely available, what remains is plain text, exactly the thing we would like to index.
- An image search would be welcome as well. This image search should be preferablely context sensitive. The images itself could stay on the remote server, but thumbnails would be nice.

I hope this document has informed you on the progress and/or motivated you to join the indexing project.