Technology Transfer in Computing Systems

D3.29: Individual TTP29 abstract

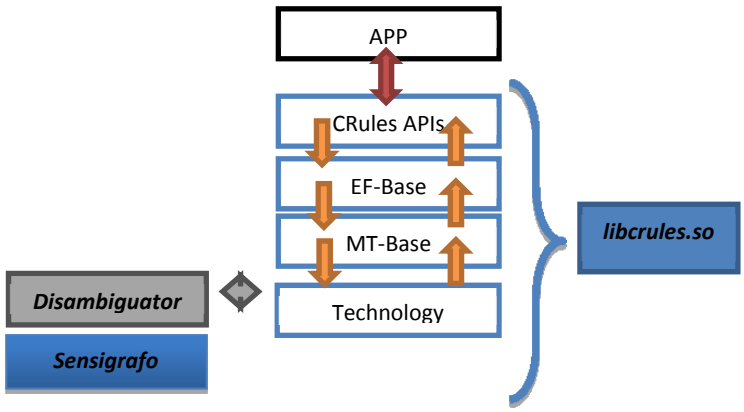
Project no.:	609491
Funding scheme:	Collaborative project
Start date of the project:	1st September 2013
Duration:	36 months
Work programme topic:	FP7-ICT-2013-10
Deliverable type:	Report
Deliverable reference number:	ICT-609491 / D3.29
WP and tasks contributing:	WP 3 / all
Due date:	29/02/2016
Actual submission date:	08/03/2016
Responsible Organization:	UNIMORE
Dissemination Level:	Public
Revision:	1.0

TETRACOM D3.29: SemBoost: order-of-magnitude performance Boost for a leading Semantic engine

Marko Bertogna, Paolo Burgio and Micaela Verrucchi (University of Modena and Reggio Emilia, Italy), Marcello Pellacani (Expert System S.P.A. Italy)

Semantic Intelligence is the ability to gather the necessary information to allow identifying, detecting and solving semantic information from a given input source such as texts or images. Similar to Operational Intelligence or Business Process Intelligence, which aims to identify, detect and then optimize business processes, semantic intelligence targets information instead of processes.

Expert System (ES) is a fast-growing SME quoted in the Italian stock market, active since 1994. Its core activity lies in the extrapolation of semantic meanings from large sets of documents. Its main product, labeled COGITO, is a semantic engine for intelligently searching selected concepts within unstructured texts, without needing to specify exactly matching keywords or sentences. COGITO has been successfully adopted within Microsoft Word suite, as "semantic spellchecker" since 1997. Another product, ADmantX, extracts and categorizes information from the text of a web page, and then it crosschecks this information with the profile of the user who is currently navigating that page, to identify the most suitable advertisement to show. Both products are based on a common layer called *Disambiguator* that is in charge of extracting core information from the web text -- the most time-consuming part of both processes. This is depicted in the following figure (extracted from the COGITO technological stack –**confidential**), where the bottom-left part is the key layer we focused on for achieving performance.



With the increasing size of documents and semantic rules to process, the search engine is subject to an increasing pressure. Current customers require faster responses to take business decisions and/or to elaborate larger sets of market data. However, the stagnating speeds of today's servers are questioning the possibility for further performance improvements. The only way-out is to exploit the immense computing power provided by modern parallel architectures.

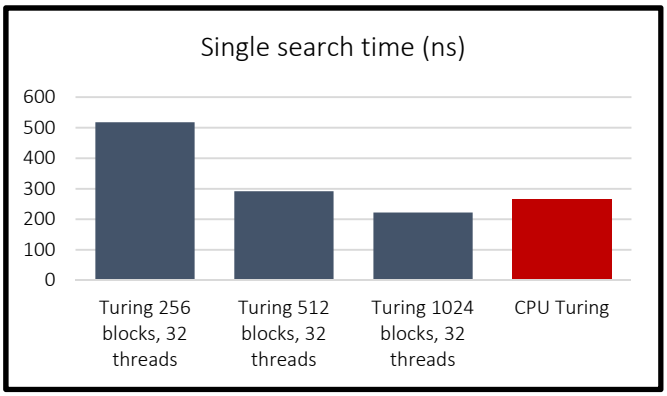
Project **SemBoost** allows boosting the performance of the core application of ES via a fine-grained parallelization of the principal routines for the exploration of the semantic data structures of the system. An initial investigation by the academic partner identified the most time-consuming procedures whose execution does not scale well for larger problem sizes (i.e., number of documents to process, size of the semantic graph, and number of search rules), with a more-than-linear increase in the complexity. After identifying the core functionalities that are more amenable to (massive) parallelization, we accelerated them on many-core accelerators such as General Purpose-GPUs using techniques developed by the academic partner for achieving a higher performance and faster response time.

The key point here is the usage of commodity GPUs offering the best price/performance ratio. It is important to note how Expert System is moving to a business model more based on cloud computing, where hardware architecture are provided by external third parties. This limits the chances of adopting e.g., next generation many-core architectures, such as STM STHORM and Kalray MPPA, which are not available on the market yet. GP-GPUs such as the NVIDIA GT900 series provide the best power/performance and price/performance tradeoff currently available on the cloud market. Finally, the project will allow ES to open up for wider market possibilities and significant savings:

- reduction of the servers cost;
- speed-up of the semantic search engine to meet further market requests for performance;
- speed-up of the internal development cost;
- Reduction of power consumption.

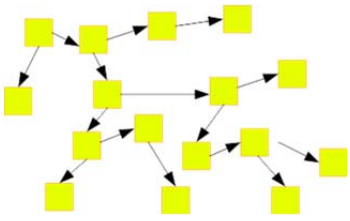
Figure below shows the performance results achieved with a first "naïve" implementation of the core functionality, a string search/matching against the Sensigrafo/database of concepts. We implemented it in CUDA on a standard consumer NVIDIA GTX980 – codename "Turing":

NVIDIA GEFORCE GTX 980 "Turing"	
Microarchitecture	Maxwell
CUDA Version	5.2
# CUDA cores	2048 in 16 SMs
Clock	1 GHz
Shared mem	96kB for 16 cores

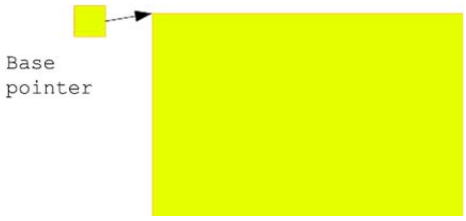


Results do not show a great performance improvement, but this implementation let us identifying the optimal tradeoff e.g., between CUDA threads/blocks. Subsequently, we applied **CompactArr**, an original memory allocator developed by us in the domain of a previous research project named LightKer, and further optimizing the code, achieving a final speedup of 8x compared to CPU-based implementation, and a 2-3x reduction in the memory footprint of the Sensigrafo.

CompactArr (CA) is an ad-hoc data structure explicitly designed for host+accelerator systems. It is based on memory offsets instead of pointers; hence it is i) compact, ii) cache-friendly and iii) easy to move across different memory spaces. It tackles the limitations of traditional pointer-based data structures, hence sparse in memory (not cache-friendly) and hard to move across different memory spaces. See figure below:



"Traditional" pointer-based data structure



CompactArr

Results are covered by NDA and are available under request, by writing to the TTP contact, prof. Marko Bertogna.