www.vineyard-h2020.eu

D2.2: Workload & Traffic Pattern Characterization

DOCUMENT ID	**D2.2.**	**CONTRACT START DATE**	**1st FEBRUARY 2016**
DUE DATE	**31/10/2016**	**CONTRACT DURATION**	**36 Months**
DELIVERY DATE	**31/10/2016**		
CLASIFICATION	**Confidential**		
AUTHOR/S	**FORTH,NEURASMUS,ICCS,NEUROCOM, ATHEX, LEANXCALE**		
DOCUMENT VERSION	**1.0**		

PARTNERS
Institute of Communications and Computer Systems, Maxeler Technologies, Bull Systems, Queen's University of Belfast, Foundation for Research and Technology, The Hartree Centre / Science and Technologies Facilities Council, Neurasmus BV, Neurocom Luxembourg, Hellenic Exchanges SA, Holding, Clearing, Settlement and Registry, LeanXcale, Loba

European Commission

Co-funded by the Horizon 2020 Framework Programme of the European Union under Grant Agreement nº 687628

1 EXECUTIVE SUMMARY

In this deliverable, we discuss the characteristics of applications that we expect will benefit from the use of accelerators. We divide applications in three categories:

- Micro-benchmarks that have been proposed and used for evaluating accelerator-based architectures. Such benchmarks cover a range of functions that are either performed today in popular domains or are projected to become important.
- Synthetic workloads, as composed from micro-benchmarks, to reflect mixed workloads in datacenters.
- VINEYARD applications, mainly brain modeling and financial, that are projected to benefit from such architectures.

We present our characterization for each category as well as the extensive infrastructure we developed and used for this purpose. From our characterization, we identify the following points:

Analysis of kernels and potential use of accelerators in the datacenter

- Execution time of kernels and tasks has a very large span, varying in the tasks we examine from tens of microseconds to a several seconds.
- Input and output data sizes can vary between a few hundred Bytes to tens of MBytes per task.
- For certain tasks, the execution time on GPUs is greatly reduced compared to the execution time of the same task on a CPU core.
- GPUs require extensive data transfers that in cases dominate execution time. Therefore, GPUs are better utilized when memory transfers are limited to a minimum, or when they execute compute intensive operations.
- Standalone execution of tasks results in small variations in tail statistics, especially task execution time. However, concurrent execution of tasks on a datacenter, regardless of their duration or the targeted accelerator, increases variability of each task's execution time by orders of magnitude, as reflected in tail statistics.

Synthetic workloads

- Workloads tend to show bursts in job arrival.
- Workloads tend to consist of an overwhelming majority of small jobs, with durations ranging between a few seconds and a few minutes, which however consume a minority of resources; a tiny percentage of huge jobs, lasts for several hours.
- Given the data from a real-life datacenter, consisting of a big cluster with hundreds of machines, it is not imminently evident how to come up with a scaled-down workload description, suitable for smaller clusters with tens of machines. We provide synthetic workload scenarios by following the insights available in recent literature and by creating short and long duration jobs by using the benchmarks available to us.

Brain simulation and neurocomputing

- The input/output footprint of jobs, as well as their execution time, depend on the network size and the connectivity density.
- Job input/output data sizes range from 60 KBytes (mouse-level) to 4MBytes (human-level), same for input and output.
- Job storage footprint is one order of magnitude larger than the input (or output), approaching 50 MBytes for humans.

PARTNERS
Institute of Communications and Computer Systems, Maxeler Technologies, Bull Systems, Queen's University of Belfast, Foundation for Research and Technology, The Hartree Centre / Science and Technologies Facilities Council, Neurasmus BV, Neurocom Luxembourg, Hellenic Exchanges SA, Holding, Clearing, Settlement and Registry, LeanXcale, Loba
European Commission | Co-funded by the Horizon 2020 Framework Programme of the European Union under Grant Agreement n° 687628

- Job execution time (C reference code on typical server) is in the order of hundreds of seconds for networks with 100K neurons (cat-level).
- Thus, for a 50 μs simulation-step, the total experiment time for a 100-millisecond simulation may last a few days.
- Profiling results show that GAP Junction computations the vast majority of the total floating-point operations, thus constituting a target that we plan to accelerate in VINEYARD. Acceleration of the applications tasks is expected to enable longer simulations (several seconds) of larger networks.

Financial applications

- Exchange market size will need to scale by one order of magnitude from about 2000 instruments to about 20000. Two types of orders (limit orders and quotes) dominate the volume of transactions and they constitute more than 99% of the served volume.
- Sustained load will need to increase by more than 3 orders of magnitude from a few hundred operations per second to about 100,000 operations/sec.
- The adoption of the FIX protocol, despite the benefits it will bring, it still requires significant processing per message in the order of several 10s of microseconds per request and will constitute a significant source of overhead.
- The pre-trade system analysis shows that there are several functions (related to risk analysis) that are heavy in terms of processing and will be a challenge to execute at the envisioned rates and in real-time.

Spark-based Analytics

- Analytics workloads and especially spark-based applications tend to be CPU-heavy.
- Spark makes good use of memory, also by reducing access to I/O devices.
- Analytics codes do not match very well modern processor architectures exhibiting low IPC and poor cache behavior.

Transactional Analytics

The analysis of our Conflict Manager subsystem in update intensive workloads shows that:

- Concurrency control is CPU bound and it is the initial bottleneck that we will try to solve by off-loading the main processor with FPGAs.
- Network bandwidth will become a second bottleneck when the capacity of processing conflict checks is multiplied by 50x/100x on the sever side.
- We expect to deploy a smaller number of physical nodes by improving the conflict management service, reducing the energy consumption (less machines) and the cost of the maintenance of the nodes (both bare-metal/cloud).

Overall, the expectation is that acceleration has an important role to play, but there are significant hurdles to be overcome.

PARTNERS
Institute of Communications and Computer Systems, Maxeler Technologies, Bull Systems, Queen's University of Belfast, Foundation for Research and Technology, The Hartree Centre / Science and Technologies Facilities Council, Neurasmus BV, Neurocom Luxembourg, Hellenic Exchanges SA, Holding, Clearing, Settlement and Registry, LeanXcale, Loba
European Commission | Co-funded by the Horizon 2020 Framework Programme of the European Union under Grant Agreement n° 687628