VINEYARD

Versatile Integrated Accelerator-based Heterogeneous Data Centers

ΕΠΙΣΕΥ ICCS · MAXELER Technologies — MAXIMUM PERFORMANCE COMPUTING · BULL · Queen's University Belfast · FORTH Institute of Computer Science · Hartree Centre Science & Technology Facilities Council · neurasmus · NEUROCOM LUXEMBOURG · ATHEX Athens Stock Exchange · LEANXCALE · LOBA
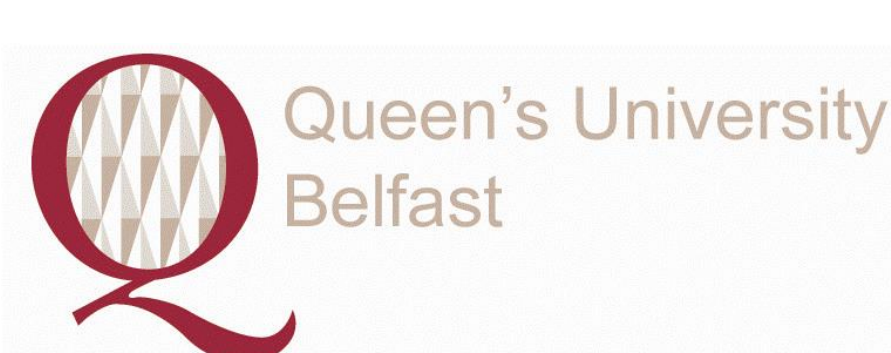
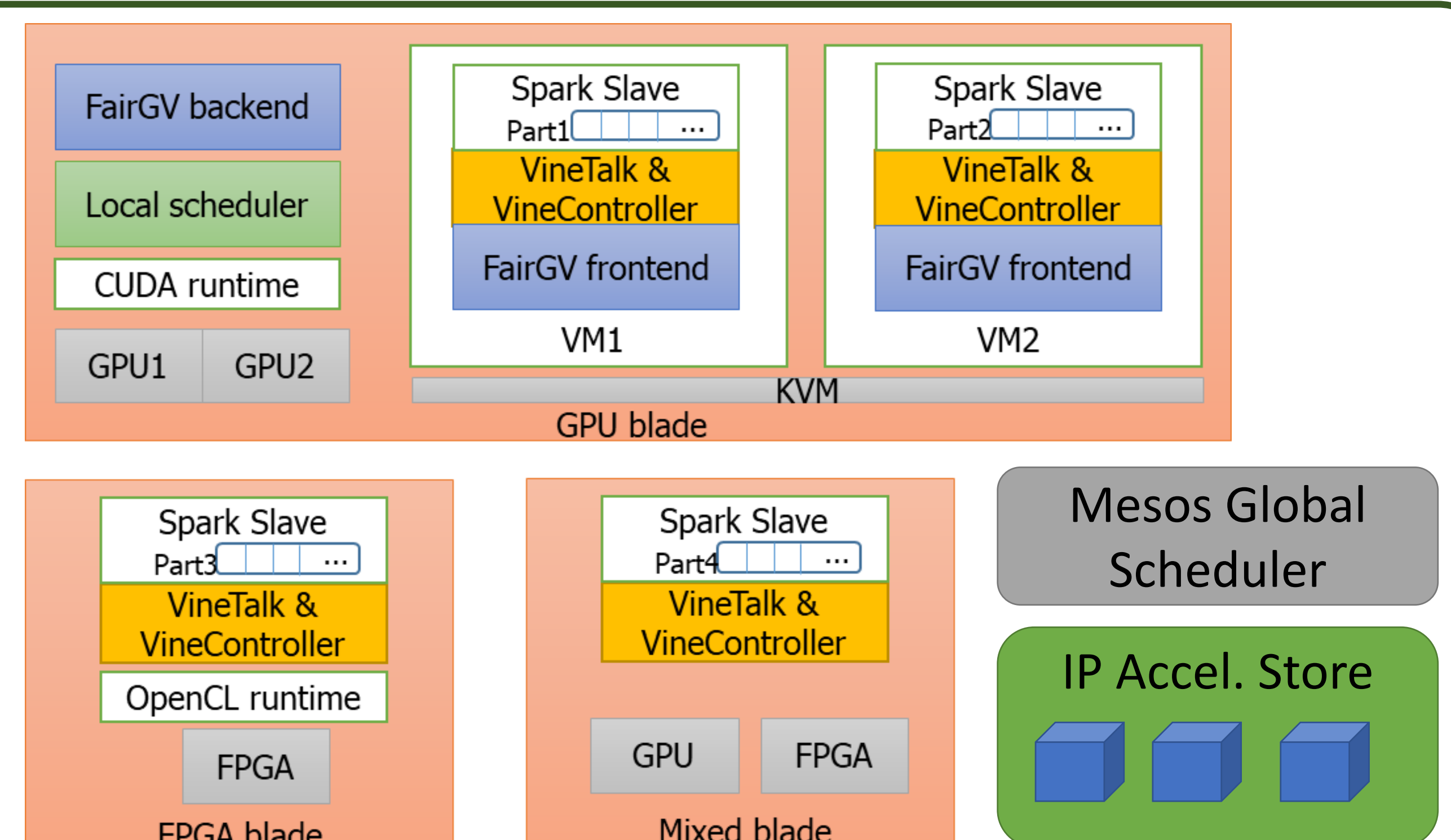## Accelerated Cloud computing, in green data centers, seamlessly…

VINEYARD aims to:
- Build energy-efficient data centres based on novel programmable **hardware accelerators** (namely Dataflow engines and FPGA-coupled servers) that can speedup cloud computing and data analytic applications.
- Develop a **high-level programming framework** for allowing end-users to seamlessly utilize these accelerators in heterogeneous computing systems by employing typical data-centre programming frameworks (i.e. Spark).

Cloud applications (ML, Data analytics):

APACHE **Spark**

Apache **MESOS**

amazon web services **EC2**

- Transparent use of FPGAs and GPUs in distributed computing systems through ready-to-use APIs for **Spark**
- Efficient resource allocation and virtualization of accelerators through **Mesos** and VineTalk
- Seamless integration with Accelerators store (e.g. **Amazon** AWS EC2 F1)

FairGV backend · Local scheduler · CUDA runtime · GPU1 · GPU2

Spark Slave Part1 — VineTalk & VineController — FairGV frontend — VM1

Spark Slave Part2 — VineTalk & VineController — FairGV frontend — VM2

KVM · GPU blade

Spark Slave Part3 — VineTalk & VineController — OpenCL runtime — FPGA — FPGA blade

Spark Slave Part4 — VineTalk & VineController — GPU — FPGA — Mixed blade

Mesos Global Scheduler

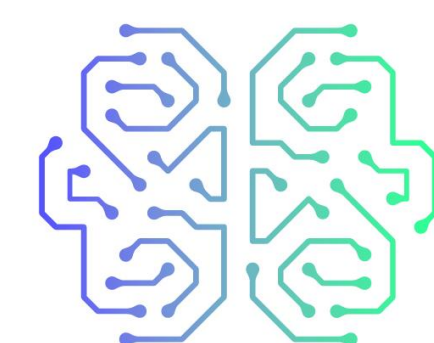IP Accel. Store

---

Transparent and integrated use of
- Maxeler Dataflow engines
- FPGAs, GPUs, and
- Intel Xeon Phi's and Intel+FPGAs for:
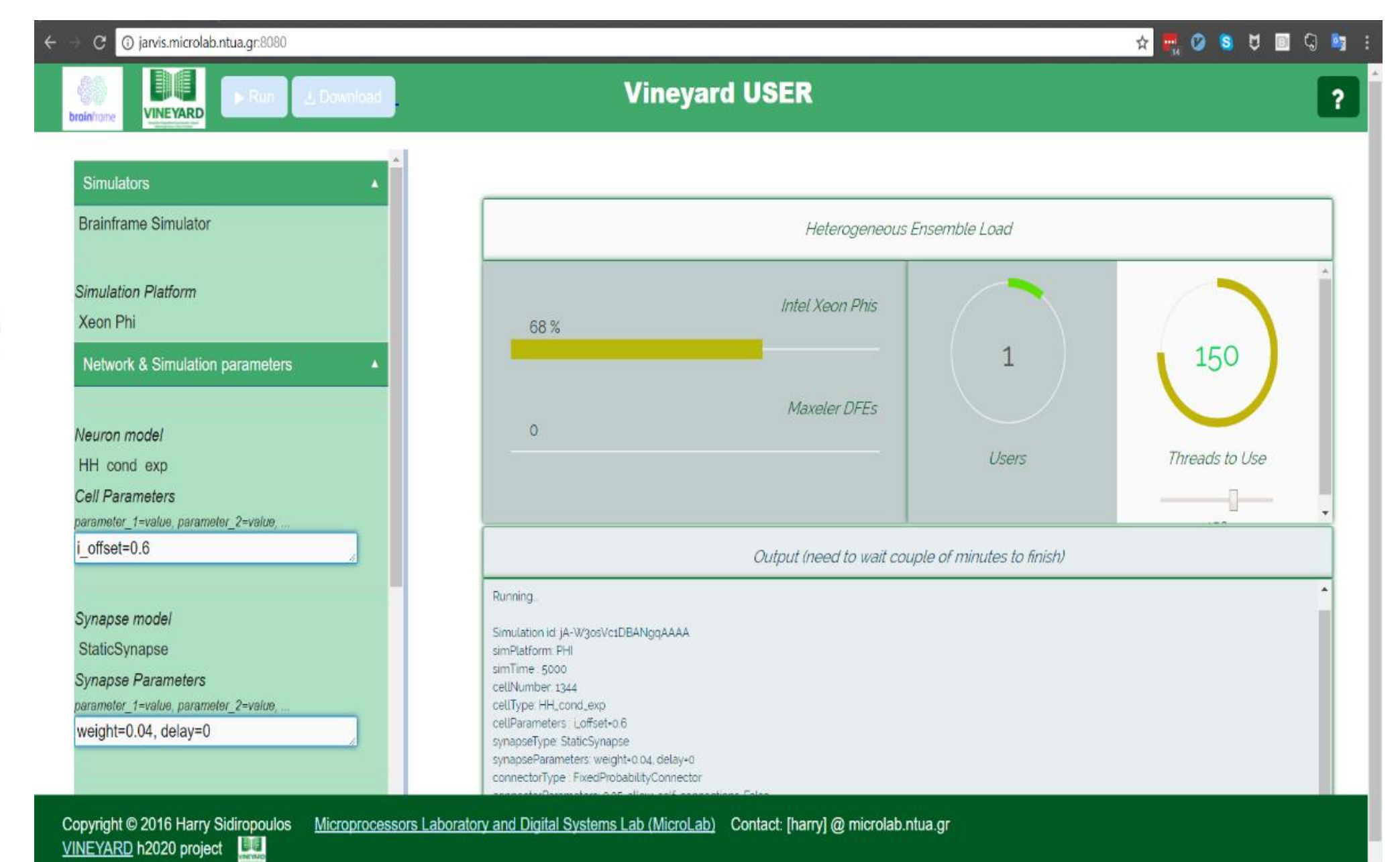
Neurocomputing applications
  brain simulations using PyNN or NeuroML (front-end) & optimally selected heterogeneous accelerators (back-end)
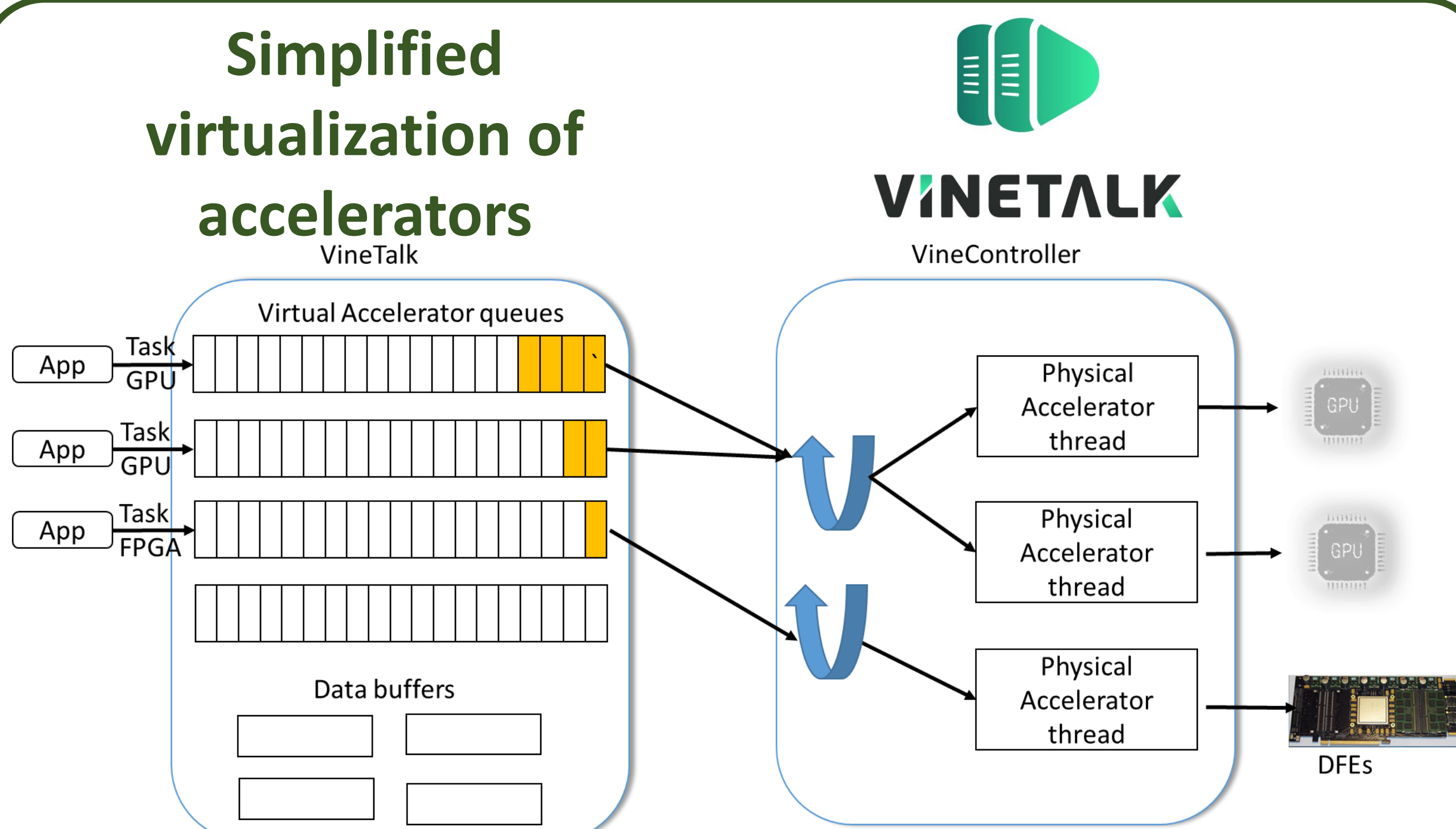
Financial applications – Risk valuation

Data management and Databases – Filtering, Sorting, Hashing

brainframe

Vineyard USER

Simplified virtualization of accelerators

VineTalk · VineController

App Task GPU · App Task GPU · App Task FPGA

Virtual Accelerator queues

Data buffers

Physical Accelerator thread · Physical Accelerator thread · Physical Accelerator thread

GPU · GPU · DFEs

Speedups and Energy Efficiency

- **25x** on Machine Learning (ML) applications
  - Logistic Regression
  - K-means
- **2x** of ML applications over Apache Spark
- **30x** for financial applications (Risk valuation – Black&Scholes, Black77, Binomial)
- **8x** for Database and Data Management
- up to **91%** energy savings over homogeneous system