Overview

VividSparks products are designed to meet the constantly changing needs of the modern Data Center, providing up to 90X performance increase over CPUs for common workloads, including AI/ML, automotive computing, GPU acceleration and analytics.

With complex algorithms evolving faster than silicon design cycles, it's impossible for fixed function GPU and CPU to keep pace. Our products provide reconfigurable acceleration that can adapt to continual algorithm optimizations, supporting any workload type while reducing overall cost of ownership.

VividSparks Products Specifications

	FalKon	Tez	RacEr	SupersoniK
Cores	16	32	512	64
Frequency	125MHz (FPGA) 1.1GHz (ASIC)	100MHz (FPGA) 1.1GHz (ASIC)	300MHz (FPGA) 1.4GHz (ASIC)	100MHz (FPGA) 1.1GHz (ASIC)
Interface	PCIe Gen3X16 (FPGA) AXI (ASIC)	PCIe Gen3X16 (FPGA) AXI (ASIC)	PCIe Gen3X16 (FPGA) AXI (ASIC)	PCIe Gen3X16 (FPGA) AXI (ASIC)
OS support	CentOS7, Ubuntu 20.04	CentOS7, Ubuntu 20.04	CentOS7, Ubuntu 20.04	CentOS7, Ubuntu 20.04
Supproted Languages	C/C++,Python/PyTorch	C/C++,Python	C/C++, PyThon, PyTorch,FORTRAN, OpenCL/GL,CUDA	C/C++,GROMACS OpenFOAM
Compiler	VividKompiler	VividKompiler	VividKompiler	VividKompiler
Number Format	POSIT	POSIT	POSIT	POSIT
DDR Total Bandwidth	77GB/s	77GB/s	77GB/s	77GB/s
DDR Total Capacity	64GB	64GB	64GB	64GB

Heighlights

Fast-Highest Performance

- Up to 90X higher performance than CPUs[1] on key workloads at 1/3 the cost[2]

- Over 4X higher inference throughput[3] and 3X latency advantage over GPU-based solutions[3]

1: BlackLynx Elasticsearch on Alveo versus EC2 c4.8xlarge

2: Based on CapEx & OpEx savings for AI/ML to HPC applications on Alveo vs dual-socket Intel Xeon Platinum servers

3: Measured on various applications against NVidia P4

Adaptable – Accelerate Any Workload

- AI/ML to any workload using the same products

- As workloads algorithms evolve, use reconfigurable hardware to adapt faster than fixed-function accelerator card product cycles

Accessible – Cloud⟷On-Premises Mobility

- Deploy solutions in the cloud or on-premises interchangeably, scalable to application requirements

- Applications available for common workloads, or build your own with application development tool

Products Architecture

VividSparks products consists of an array of tiles, connected by a 2-D mesh network called the manycore accelerator, with an attached external memory and I/O system. Most tiles contain processing, memory, and communication routers. Processing in a tile is done with CPU cores and specialized accelerator cores The accelerator cores are added to personalize the RacEr architecture and to improve energy/performance for targeted applications. In addition to these tiles, the architecture features victim cache tiles, often located on the edge of the tiled array and termed column caches ($), but potentially located at other positions in the array. These victim cache tiles are in turn connected to memory controllers that interface to multiple parallel memory channels that go to DRAM -- high bandwidth memory (HBM), DDR4, or other. The Figure 1 on next page shows the high-level view. In some cases, some of the CPU cores might be replaced with accelerator tiles.

Each CPU core contains a 4KB direct-mapped instruction cache (1024 instructions), and a 4KB local data memory. The cores features non-blocking loads and stores, which allow them to overlap the memory latency to remote memories in the system while they execute non-dependent instructions. These word-level accesses go out onto the 2-D mesh network to the remote tile, cache or dram that owns the address in question.

Within the architecture, we have the concept of a tile group, which is a physically contiguous subarray of tiles. Tile groups work together to perform cooperative multiprocessing, where a group of cores shares a set of banked memories and distributes shared data structures across these banks. The cores use a bulk synchronous programming model, where a program is divided into phases in which each address is either read/write-owned by a single tile, read-owned by everybody, or requires atomic operation/mutex enabled atomic accesses. Between each phase, the tiles synchronize via a synchronization barrier. This allows a group of tiles to bring in a chunk of memory from DRAM, operate on that data in parallel with high locality, and then write it back to DRAM. Access to data structures within a tile group enjoys reduced energy usage, latency and increased throughput relative to access to data structures in DRAM, or in the column caches.

Qualified Servers

A list of servers on which VividSparks products are fully qualified can be found here: https://www.xilinx.com/products/boards-and-kits/alveo/qualified-servers.html.

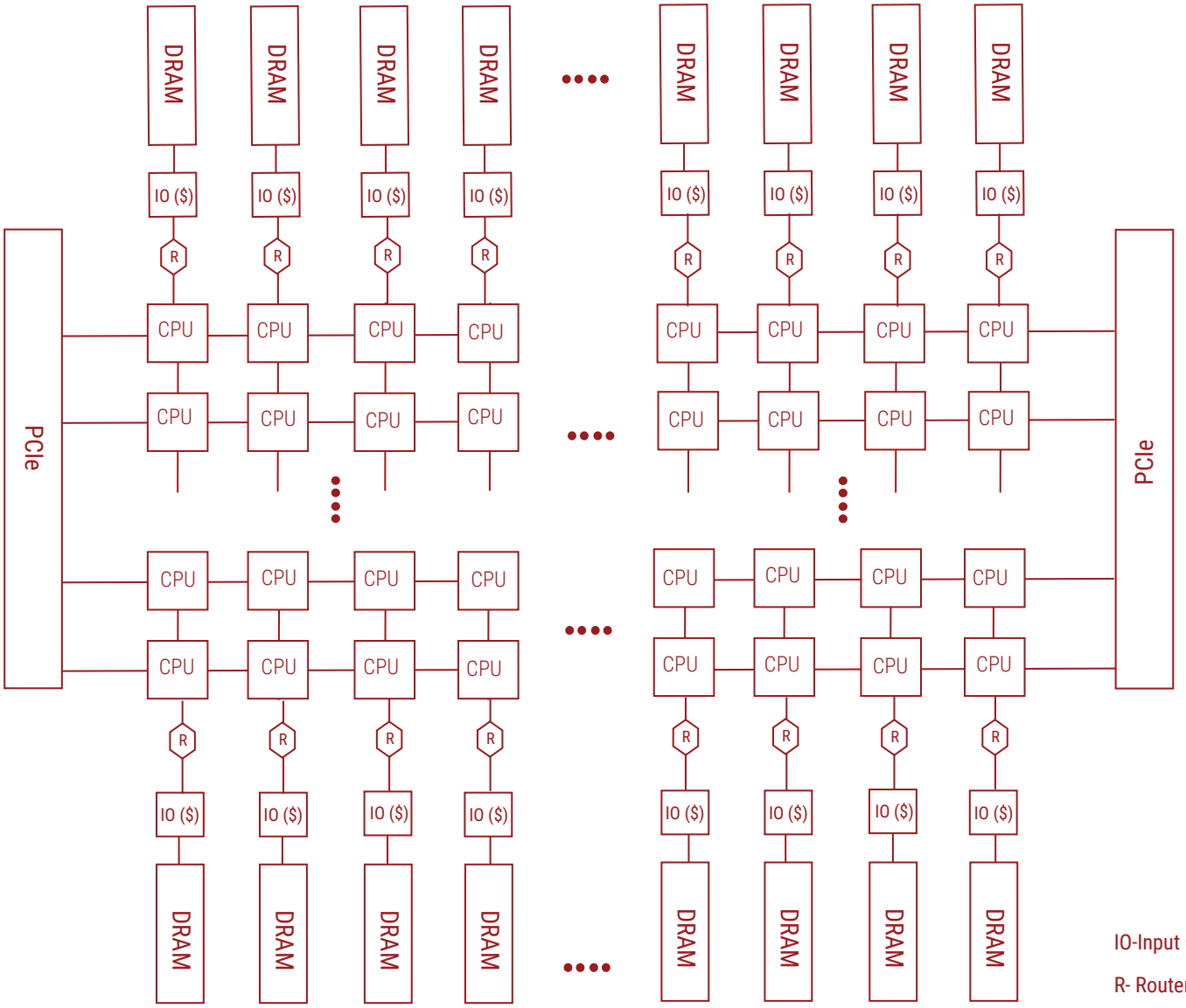

Figure 1: Architecture

Benchmarks

A list of benchmarks can be found here: https://vividsparks.us and then click on RESOURCES--> BENCHMAKRS

VividSparks IT Solutions
#38 BSK Layout, Hubli-580031, India.
https://vividsparks.tech
info@vividsparks.tech