https://www.argo-osr.org/handouts/

Future Hardware

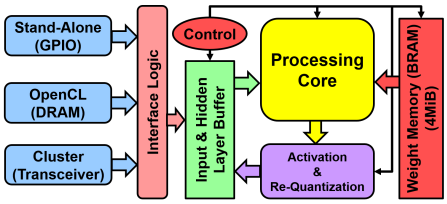
Pete Beckman (*PI*), Kamil Iskra, Swann Perarnau, Kazutomo Yoshii

Shifting paradigms in HPC systems' designs put pressure on the system software. We are continuously exploring emerging new hardware trends and devices, looking for how best to exploit the new capabilities in both traditional HPC workloads and emerging ones such as machine learning. We try to identify use cases that vendors may not have thought of and to figure out how the new features should be integrated into existing runtime systems and operating systems and in what way such features should be exposed to applications.

A significant portion of this work is covered by NDAs so it cannot be discussed here. Below we present a small sample of what *can* be made public.

FPGAs

TRIP (An Ultra-Low Latency, **T**eraOps/s **R**econfigurable **I**nference **P**rocessor for Multi-Layer Perceptrons) is a "soft" deep neural network for MLPs. It features a hybrid OpenCL-Verilog design, where the MLP engine is written in Verilog and OpenCL is used for data exchange between the CPU and the FPGA. It runs on the Nallatech 385A platform (Arria10).



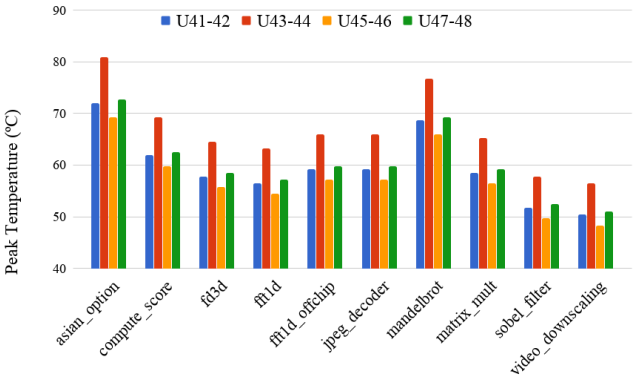
The table on the right shows how such a targeted, custom design can outperform more general-purpose vendor offerings. We continue improving the design with features such as mixed precision, more efficient data movement, and a scalable OS/runtime. This project is a collaboration with Boston University.

Table 2: ECP-Candle Performance Comparison for Single Input Vector

Inference Architecture	M,N	Useful Op (%)	Performance (TeraOps/s)	Speedup
NVIDIA K80	-	-	0.02	1x
TPU	256,256	79	0.05	2.5x
TRIP Arria 10 CoProc	256,16	91	1.5	75x
TRIP Arria 10 Cluster	256,32	89	3.0	150x
TRIP Stratix 10 CoProc	256,86	88	15.5	775x
TRIP Stratix 10 Cluster	256,102	86	18.0	900x

Thermal-Aware Computing

Cooling power is a non-negligible issue in HPC systems, and thermal variations at every level (environment, hardware from transistors to clusters) make it worse. The figure on the right shows how running the same code on four nodes of the same cluster can result in significantly different temperatures of the compute elements. We are studying CPUs, GPUs, and FPGAs and are using machine-learning prediction models for better task scheduling on the system. This project is a collaboration with Northwestern University.



Contact

For more information, please contact Kazutomo Yoshii kazutomo@mcs.anl.gov or Pete Beckman beckman@mcs.anl.gov.

U.S. DEPARTMENT OF ENERGY | Office of Science
National Nuclear Security Administration
Argonne NATIONAL LABORATORY
Lawrence Livermore National Laboratory