

Software and hardware must be optimized for diverse platforms from IoT to supercomputers

Data centers

User front-end (cloud, GRID, supercomputer, etc)

Existing frameworks / algorithms

Single node

such as RPi

Algorithm / source code

Available libraries / skeletons

Compilers

Binary or byte code

Inputs

Various models

Run-time environment

Run-time state of the system

Hardware, simulators

Microsoft Azure, AWS, Google Cloud, XSEDE, PRACE, Watson...

TensorFlow, Caffe, Torch, MXNet, TensorRT, CNTK, OpenCV ...

CUDA, MPI, OpenMP, TBB, OpenCL, StarPU, OmpSs ...

C, C++, Fortran, Java, Python, byte code, assembler ...

cuBLAS, BLAS, MAGMA, ViennaCL, CLBlast, cuDNN, openBLAS, cBLAS, libDNN, tinyDNN, ARM compute lib, libxsmm, skeletons

LLVM, GCC, ICC, Rose, PGI, Lift, functional programming ...

diverse hardware: heterogeneous, out-of-order, caches (x86, ARM, CUDA, Mali, Adreno, Power, TPU, FPGA, MIPS, AVX, neon)

Linux (CentOS, Ubuntu, RedHat, SUSE, Debian), Android, Windows, BSD, iOS, MacOS ...

Users need efficient stack while balancing speed, accuracy, energy, resource usage and other costs