Existing frameworks / algorithms CUDA, MPI, OpenMP, TBB, OpenCL, StarPU, OmpSs ... Single node C,C++,Fortran,Java,Python,byte code, assembler ... Algorithm / source code **Available libraries** LLVM,GCC,ICC,Rose,PGI, (hundreds of optimizations) ... **Compilers** cuBLAS, BLAS, MAGMA, ViennaCL, CLBlast, cuDNN, openBLAS, Binary or byte code clBLAS, libDNN, tinyDNN,ARM compute lib, libxsmm,gemmlowp **Inputs** Various models diverse hardware: heterogeneous, out-of-order, caches (x86,ARM,CUDA,Mali,Adreno,Power,TPU,FPGA,MIPS,AVX,neon) **Run-time environment Run-time state** Hardware, Linux (CentOS, Ubuntu, RedHat, SUSE, Debian), Android, of the system simulators Windows, BSD, iOS, MacOS ... Users need efficient solutions to balance speed, accuracy, energy, resource usage and other costs

Diverse SW must be optimized for diverse HW from IoT and supercomputers

Data centers

User front-end (cloud, GRID,

supercomputer, etc)

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

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