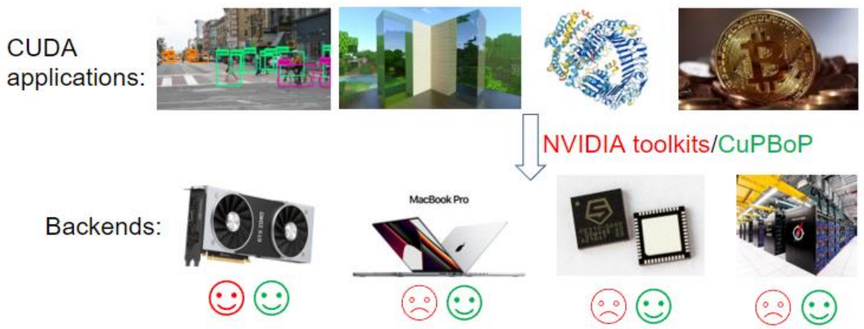
CuPBoP: CUDA for Parallelized and Broad-range Processors

Authors: Ruobing Han*, Jaewon Lee*, Jun Chen*, Bhanu Garg*, Jeffrey Young*, Mark Ahn*,
Xuele Zhou*, John Lu*, Haotian Sheng*, Blaise Tine*, Jaewoong Sim+, Hyesoon Kim*
Email: hanruobing@gatech.edu
Affiliations: *: Georgia Institute of Technology +: Seoul National University

Georgia Tech College of Computing
Center for Research into Novel Computing Hierarchies

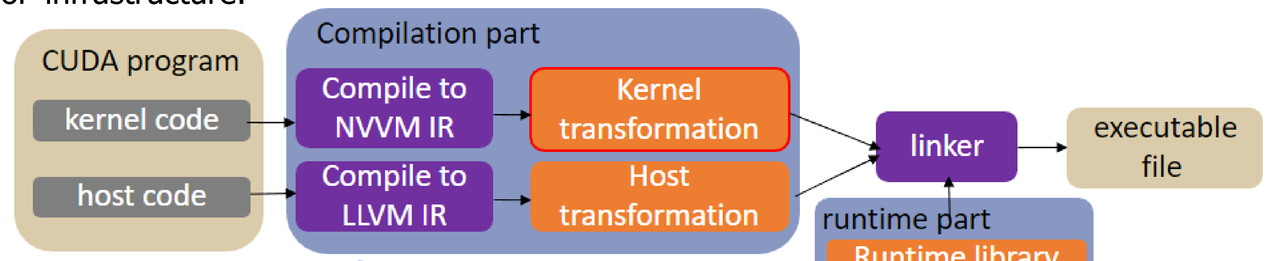
Motivation:



Relation to FORZA project:

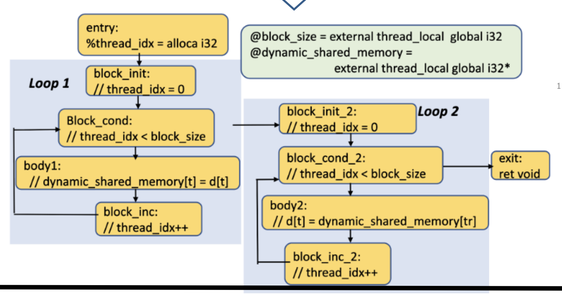
Compiler IL level conversion from CUDA applications to different computing platforms (X86 GPU, AMD GPU, Intel GPU, RISC-V GPU) including AGILE platforms.
Input program: CUDA
Translation : LLVM IR level translation

CuPBoP infrastructure:

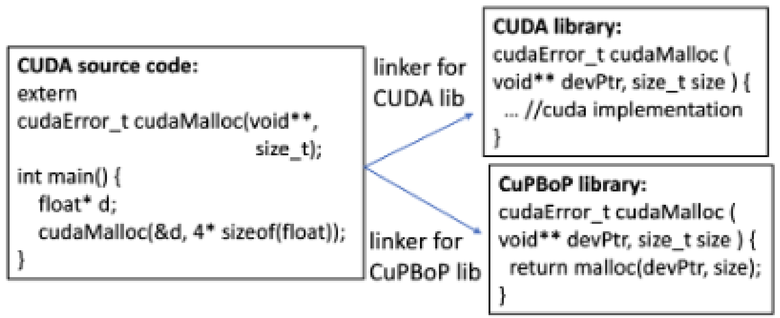


Compilation part:
Apply transformations on CUDA source code.

```
__global__ void dynamicReverse(int *d, int n) {
  extern __shared__ int s[];
  int t = threadIdx.x;
  int tr = n - t - 1;
  s[t] = d[t];
  __syncthreads();
  d[t] = s[tr];
}
```



Runtime library:
Implement CUDA APIs for non-NVIDIA devices.

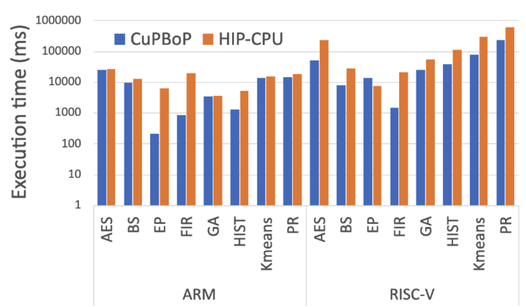


Experiment: CUDA on CPUs

Setup:
hardware: x86, AArch64, RISC-V CPUs
Benchmark: Rodinia, Hetero-mark, Crystal
Baseline: DPC++ (CUDA on Intel CPUs), HIP (CUDA on CPUs)

Framework	Compilation requirement	Runtime requirement	ISA support
DPC++	DPC++	DPC++	x86
HIP-CPU	C++17	TBB(>=2020.1-2), pthreads	x86 AArch64 RISC-V
CuPBoP	LLVM	pthreads	x86 AArch64 RISC-V

The highest hardware coverage.



The highest runtime performance.

Name	DPC++	HIP-CPU	CuPBoP	features
b+tree	correct	unsupport	correct	extern C
backprop	correct	unsupport	correct	extern C
bfs	incorrect	correct	correct	
gaussian	correct	correct	correct	
hotspot	incorrect	correct	correct	
hotspot3D	incorrect	correct	correct	
huffman	correct	unsupport	correct	extern shared memory definition
lud	correct	correct	correct	
myocyte	correct	correct	correct	
nn	correct	correct	correct	
nw	correct	correct	correct	
particlefilter	incorrect	correct	correct	
pathfinder	correct	correct	correct	
srad	correct	correct	correct	
streamcluster	correct	correct	correct	
dwt2d	segfault	unsupport	unsupport	shared memory for structure
hybridsort	unsupport	unsupport	unsupport	Texture
kmeans	unsupport	unsupport	unsupport	Texture
lavaMD	correct	correct	correct	
leukocyte	unsupport	unsupport	unsupport	Texture
mummergpu	unsupport	unsupport	unsupport	Texture
cfd	correct	unsupport	correct	cuGetErrorName
heartwall	incorrect	unsupport	incorrect	complex template
Rodinia coverage	56.5	56.5	73.9	
q11, q12, q13	unsupport	unsupport	support	warp shuffle
q21, q22, q23	unsupport	support	support	atomicCAS
q31, q32, q33, q34	unsupport	support	support	atomicCAS
q41, q42, q43	unsupport	support	support	atomicCAS
Crystal coverage	0	76.9	100	

The highest application coverage

References:

Github link
CuPBoP: CUDA for Parallelized and Broad-range Processors
COX: Exposing CUDA Warp-level Functions to CPUs