Cache blocking with CUTLASS

High Performance Scientific Computing 2018
Rio Yokota

cpu_matmult

How to compile:

```
> g++ -O3 00_base.cpp
> g++ -O3 01_loop_order.cpp
> g++ -O3 -fopenmp 02_openmp.cpp
> g++ -O3 -fopenmp -mavx -mfma 03_avx.cpp
> g++ -O3 -fopenmp -mavx -mfma 04_papi.cpp -lpapi
> g++ -O3 -fopenmp 05_sgemm.cpp -lblas
```

How to run:

```
> ./a.out 1024
```

gpu_matmult

How to compile:

```
> nvcc -Xcompiler "-O3 -fopenmp" 00_base.cu
> nvcc -Xcompiler "-O3 -fopenmp" 01_block.cu
> nvcc -Xcompiler "-O3 -fopenmp" 02_grid.cu
> nvcc -Xcompiler "-O3 -fopenmp" 03_shared.cpp
> nvcc -Xcompiler "-O3 -fopenmp" 04_sgemm.cu -lcublas
```

How to run:

```
> ./a.out 1024
```

00_base.cu can only run up to 1024

01_block.cu, 02_register.cu, 03_shared.cu, 04_sgemm.cu
can only run multiples of M

CUTLASS

```
git clone https://github.com/NVIDIA/cutlass

cd cutlass/cutlass_test

export PATH=/apps/t3/sles12sp2/cuda/9.0.176/bin:$PATH

make sgemm sm=60 transpose=nn verbose=1 keep=0

qrsh -g tga-hpc-lecture -l q_node=1 -l h_rt=0:01:00

export LD_LIBRARY_PATH=/apps/t3/sles12sp2/cuda/
9.0.176/lib64:$LD_LIBRARY_PATH

bin/sgemm_nn_sm60_nvcc_9.0
```