CONNECTION INSTRUCTIONS

- Navigate to nvlabs.qwiklab.com
- Login or create a new account
- Select the "Instructor-Led Hands-on Labs" class
- Find the lab called "Optimizing CUDA Application Performance..." and click Start
- After a short wait, lab instance connection information will be shown
- Please ask Lab Assistants for help!

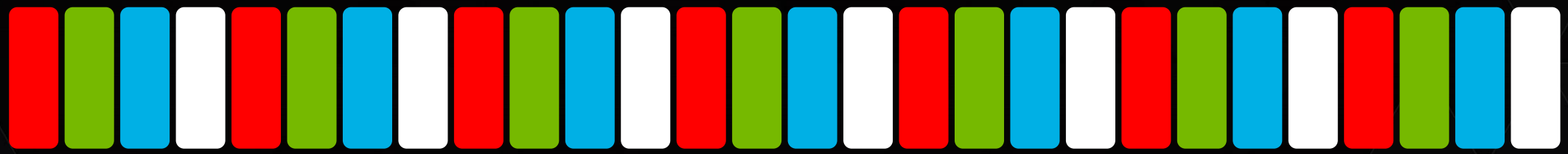
GPU TECHNOLOGY CONFERENCE

OPTIMIZING CUDA APPLICATION PERFORMANCE WITH NVIDIA'S VISUAL PROFILER

YU ZHOU (NVIDIA)

MAYANK KAUSHIK (NVIDIA)

1-D STENCIL KERNEL



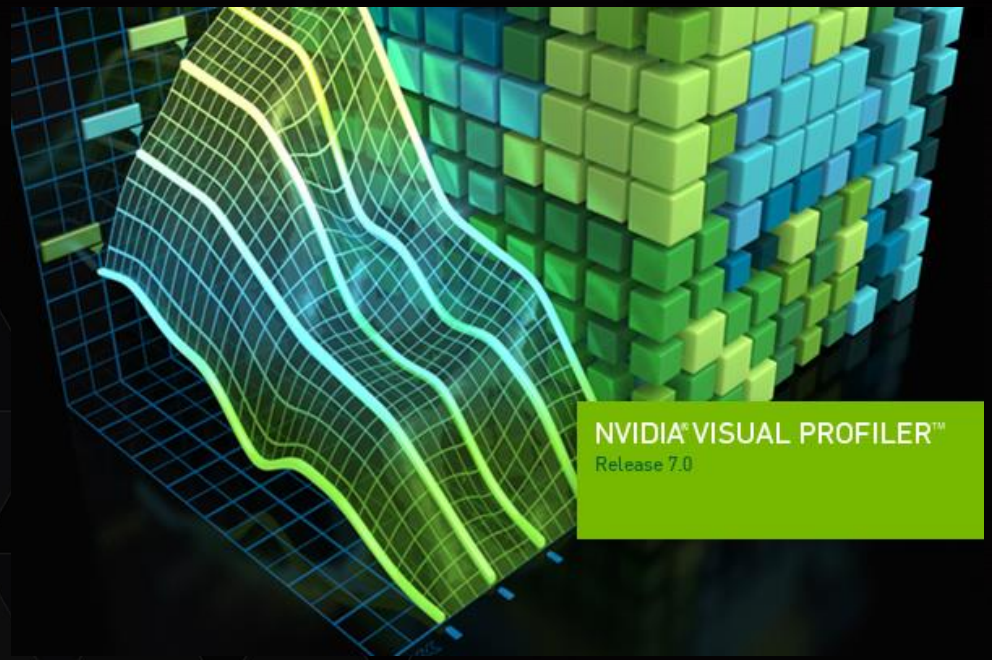
```
// Executes for each pixel
__global__ void stencilKernel(...) {

  ...

  foreach adjacent pixels:
    foreach color channels:
      out[index] += in[index + radius, channel] * weight[radius];

}

cudaMemcpy(..., in, SIZE, H2D);
stencilKernel<<< ceil(#pixels/BLOCK_SIZE), BLOCK_SIZE >>>(...);
cudaMemcpy(out, ..., SIZE, D2H);
```


NVIDIA® VISUAL PROFILER™
Release 7.0

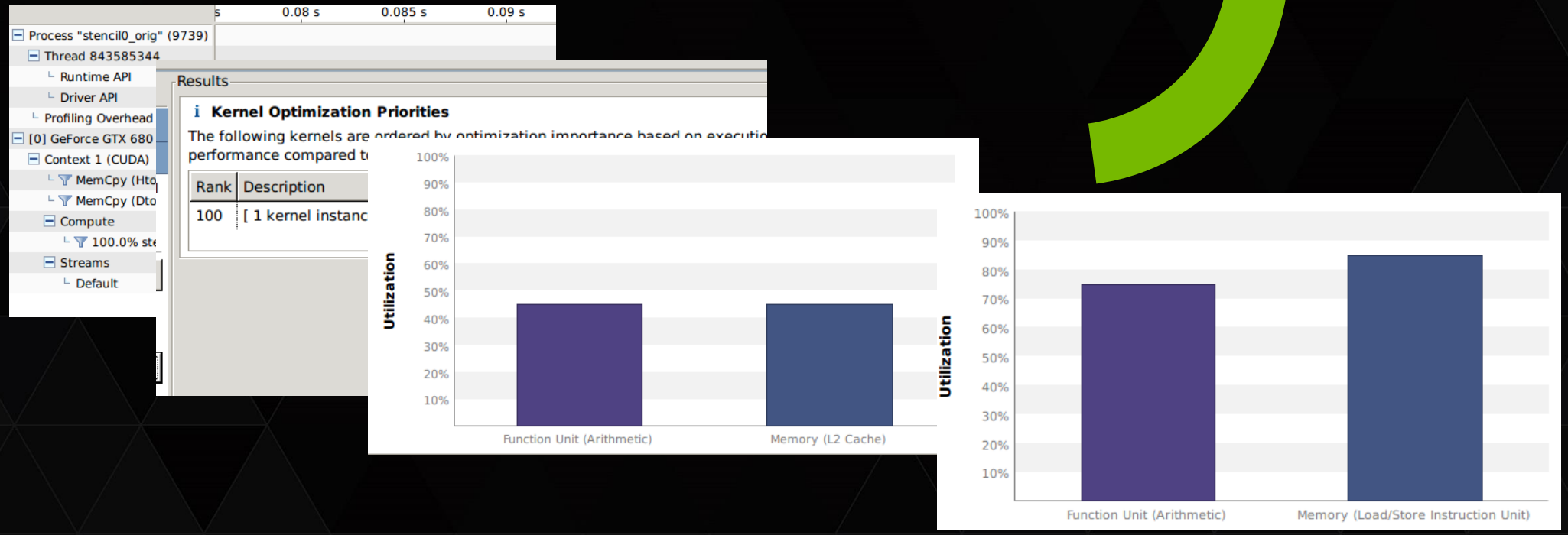

NVIDIA® VISUAL PROFILER™

HOW TO COMPILE/RUN

- **cd ~/gtc2015**
- **make stepX** (X=0,1,..,5) → stencilX_* executable
- Modify “stencilX_*.cu”
- **make clean** to restore
- ~/gtc2015/instructions.pdf

HOW TO PROFILE

- **Visual Profiler** shortcut on Desktop
- Iterative approach

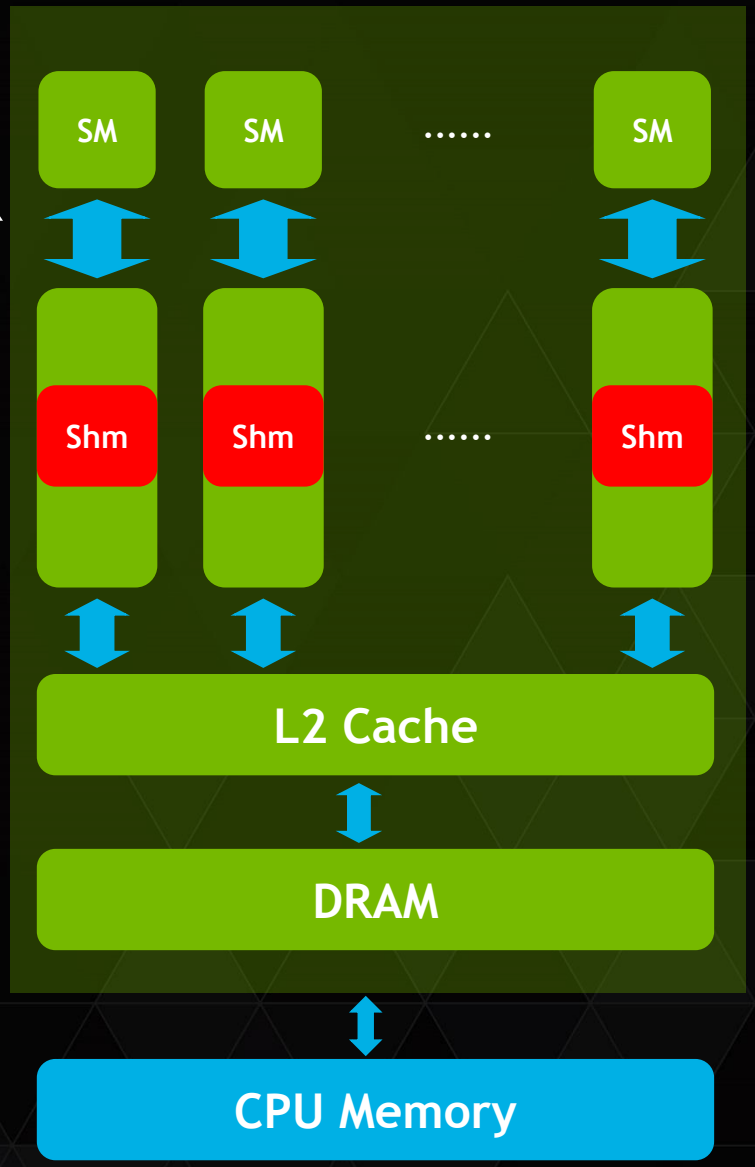


STEP1: OCCUPANCY

- Does GPU have **enough work** to do?
- Limiting factors
  - Shared memory usage
  - Register usage
  - Kernel dimensions

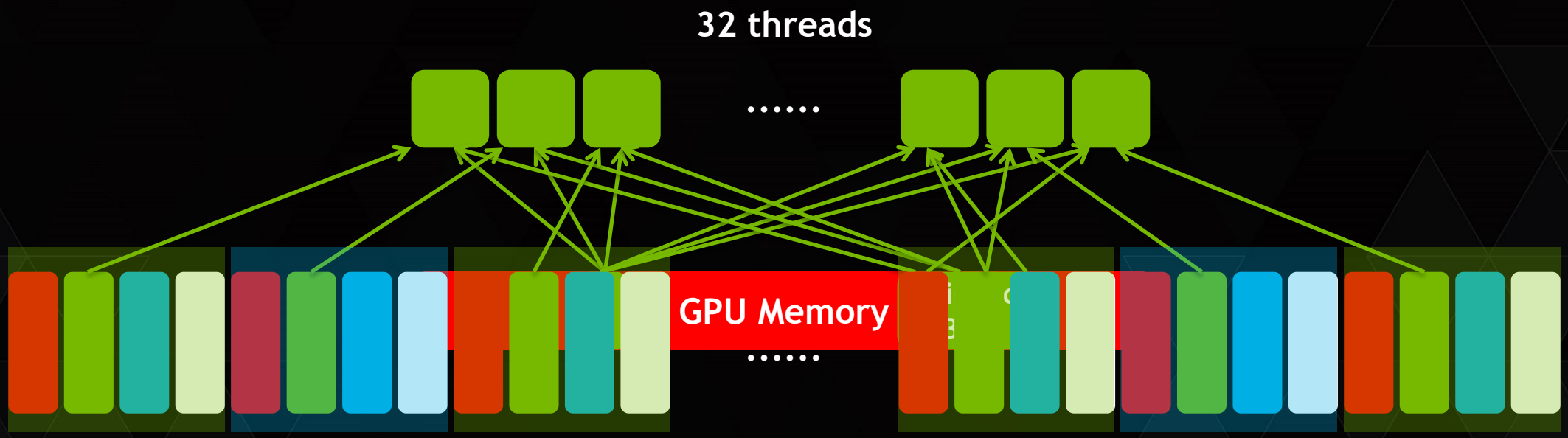
STEP2: MEMORY TRANSFER

- Transfer amount
- Duplicated transfer?
- Shared memory as **controlled cache**



STEP3: ACCESS PATTERN

- Best performance when coalesced!

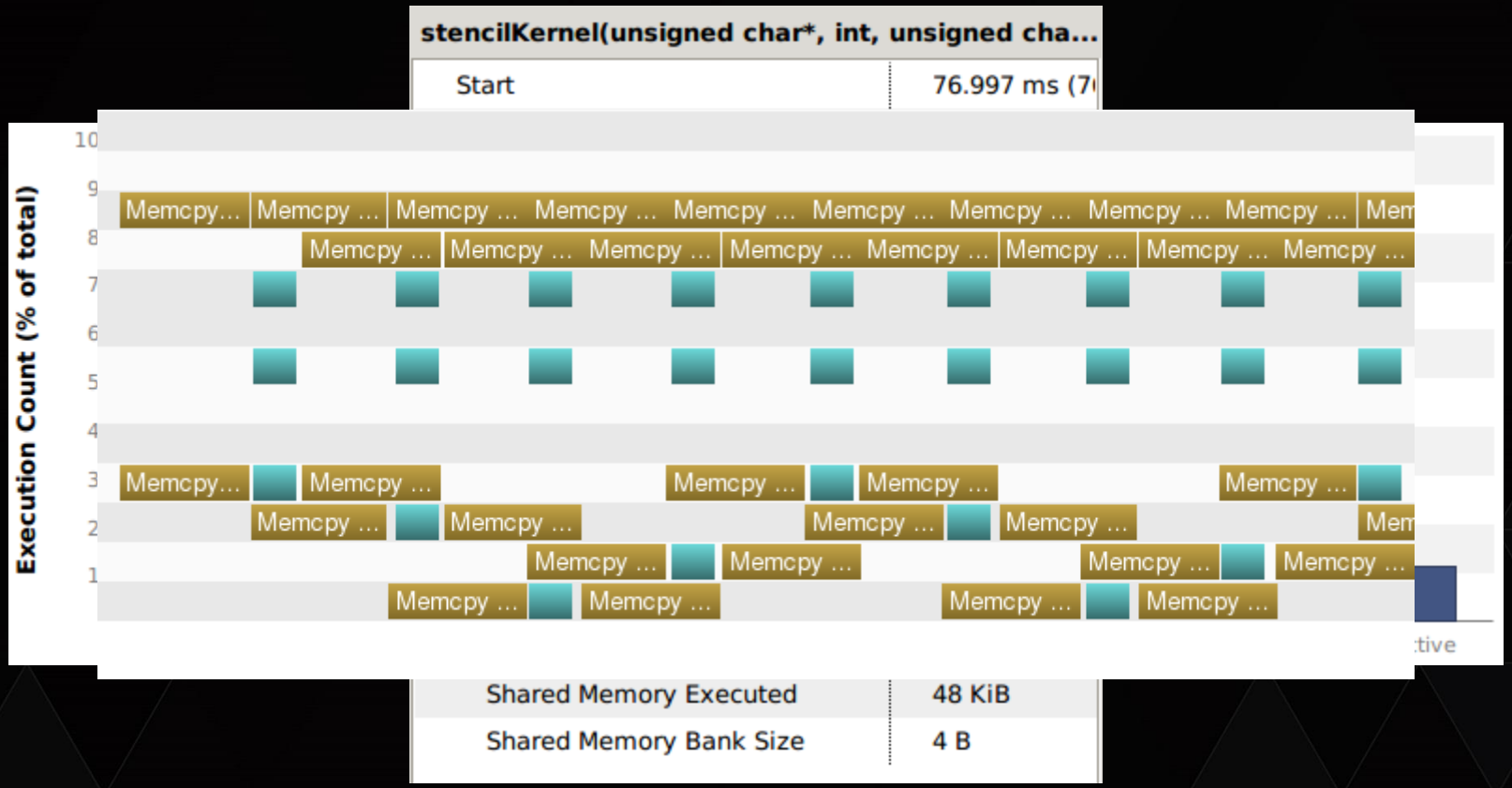


STEP4: COMPUTE UNITS UTILIZATION

- Use **kernel profile** to check hot spots
- **Balance** load between units

STEP5: BACK TO TIMELINE

- Pay attention to **application level concurrency**
- Check available features on the GPU


stencilKernel(unsigned char*, int, unsigned cha...
Start
76.997 ms (7
Execution Count (% of total)
10
9
8
7
6
5
4
3
2
1
Memcpy...
Memcpy ...
Shared Memory Executed
48 KiB
Shared Memory Bank Size
4 B

WHAT'S NEXT?

- Download today!
  Search "**download cuda**"
  cudatools@nvidia.com
- S5174 - CUDA Optimization with NVIDIA Nsight Visual Studio Edition
  15:30 – 16:50, Room 210G
- S5655 - Hands-on Lab: CUDA Application Development Life Cycle
  Thu, 14:00 – 15:20, Room 211A
- Last year's sessions
  Search "GTC on demand"
- https://github.com/yzhou61/profiler_hands_on_gtc15

GPU TECHNOLOGY CONFERENCE

THANK YOU

JOIN THE CONVERSATION

#GTC15