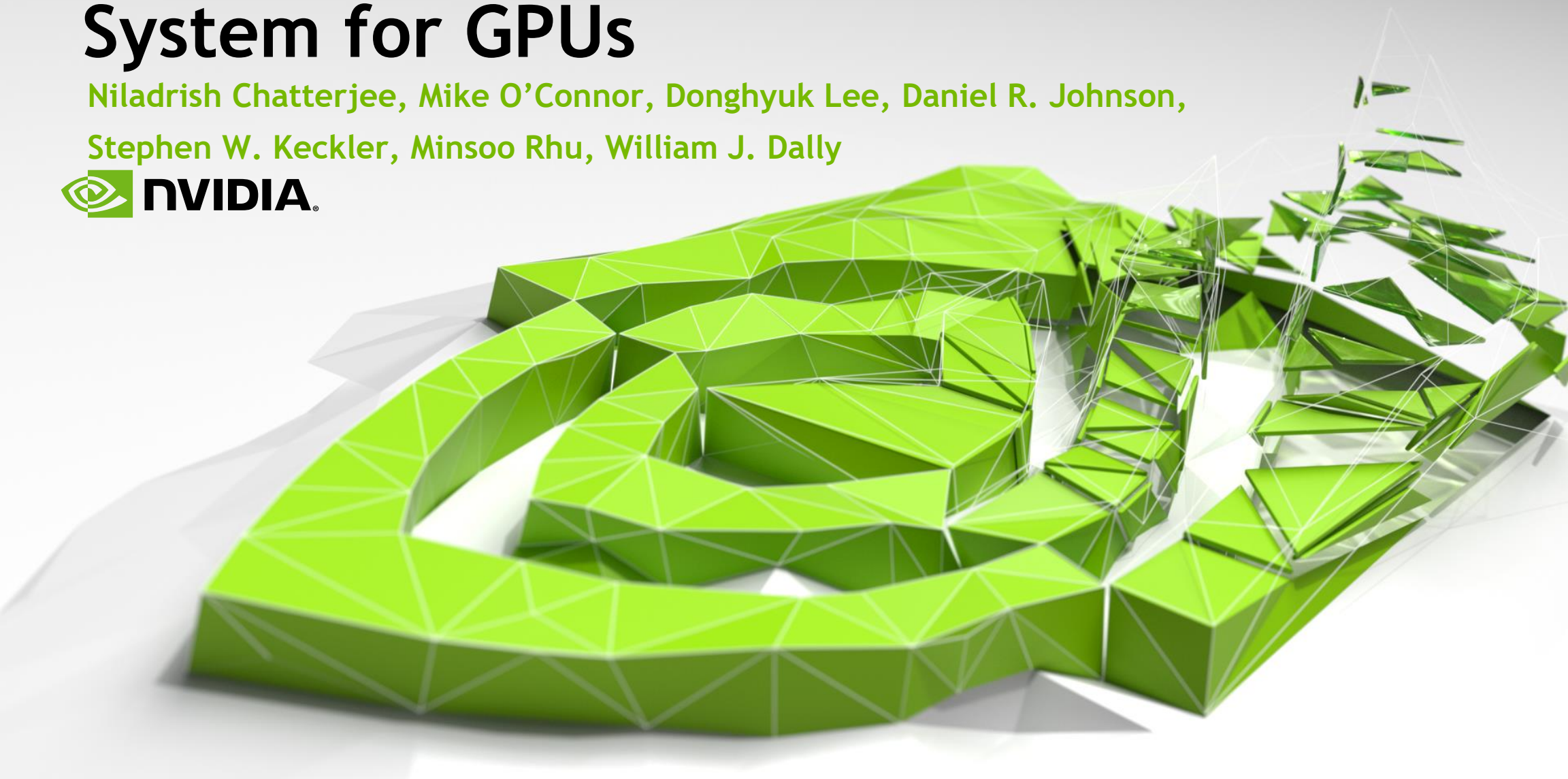
Architecting an Energy-Efficient DRAM System for GPUs

Niladrish Chatterjee, Mike O'Connor, Donghyuk Lee, Daniel R. Johnson,
Stephen W. Keckler, Minsoo Rhu, William J. Dally

NVIDIA



The Problem

GPU DRAMs consume too much energy

- 4TB/s Exascale GPU with HBM2

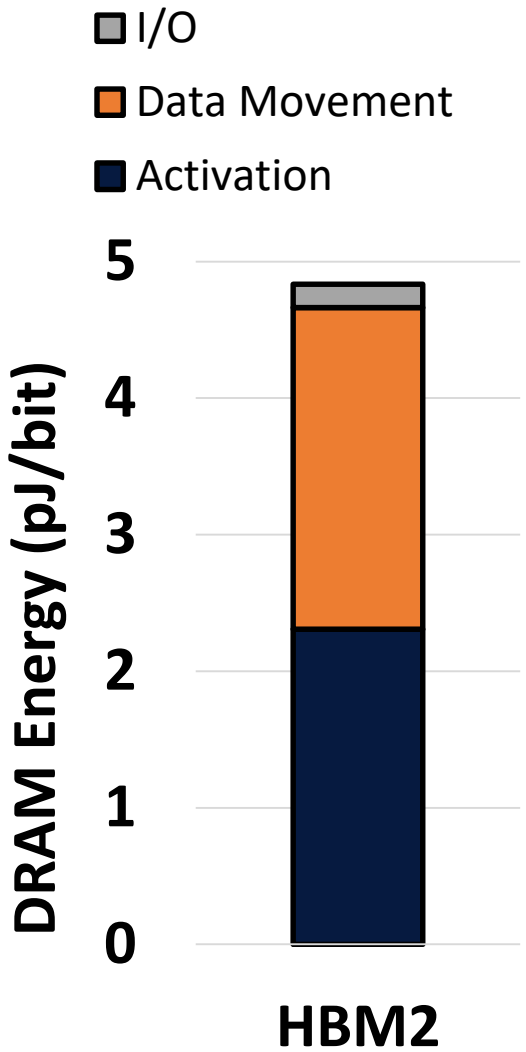
  ➡ **150W for memory !**

- Low row-buffer utilization

  ➡ **high row activation energy**

- Banks with small rows

  ➡ **high area cost**



Solution: Subchannels

Reduced effective row-size

- Memory controller directed activation granularity
- Wide datapath split into parallel narrow subchannels

35% DRAM energy savings

13% GPU performance improvement

2.6% DRAM die-area overhead