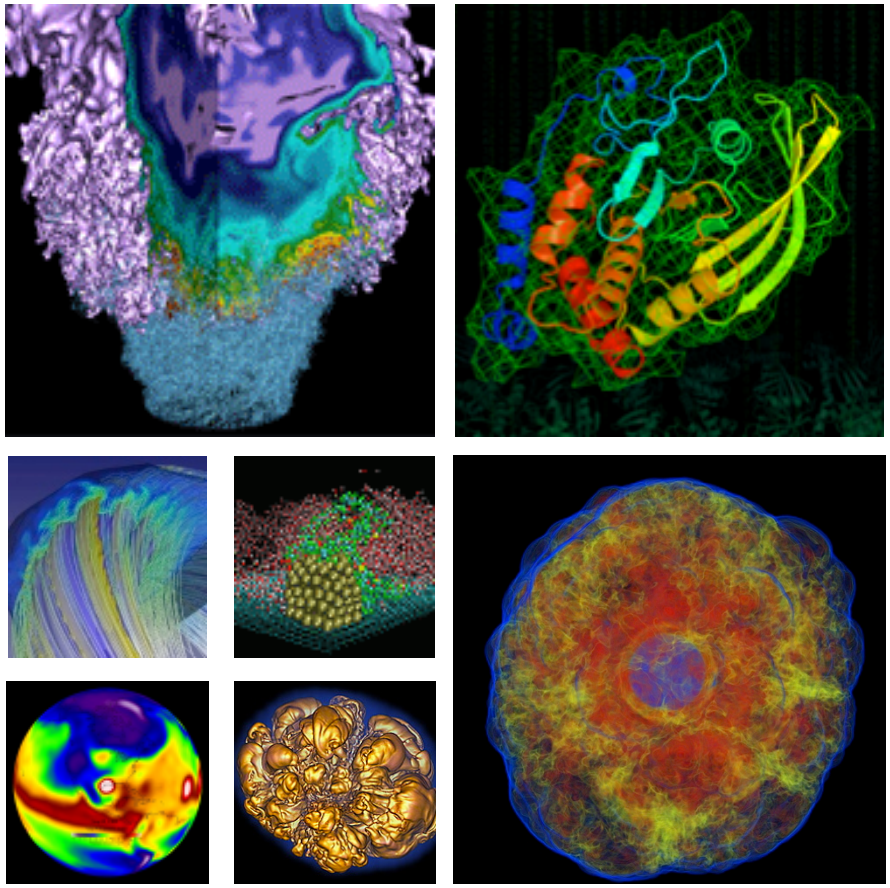
Automatic deep understanding of deep learning codes performance in many-core processors



NERSC

Tareq Malas
Thorsten Kurth
Jack Deslippe

November 16, 2016
IXPUG BoF Supercomputing'16

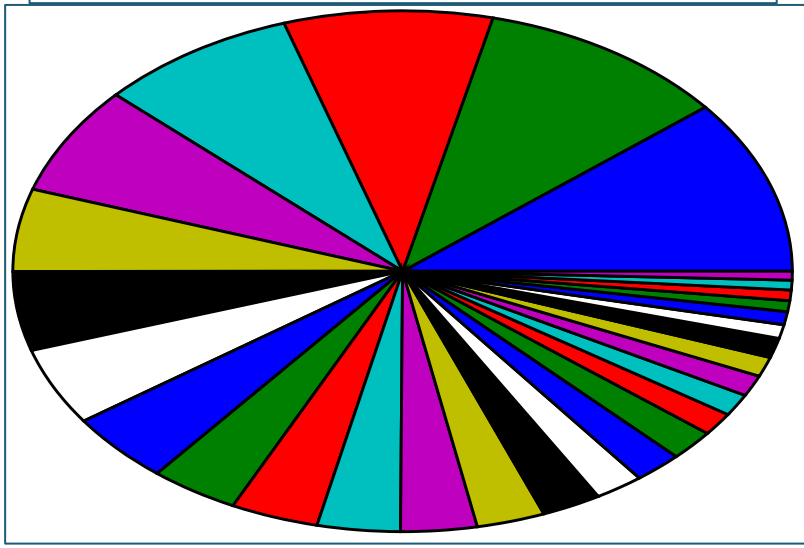
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

Deep Neural Networks and HPC

- **DNN gained significant importance in recent years**
- **Increasing computational demands**
- **Flat time distribution among many kernels**
  - With varying DNN kernels across applications
- **Automating the DNN's performance understanding is a necessity**

Alexnet Time breakdown in KNL



Approach: Roofline per kernel

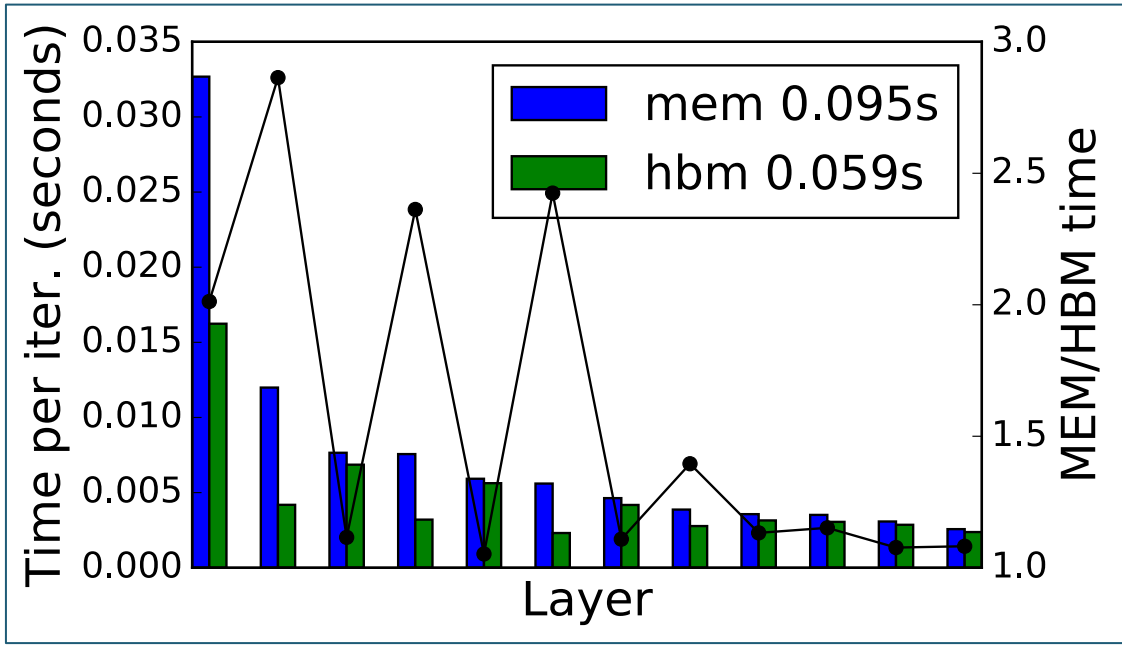
- **The Roofline model is an excellent tool to understand the compute/memory bottlenecks**
- **We automatically collect Roofline data per layer in the famous Caffe DNN framework**
  - **Memory transfers** are obtained using LIKWID performance tools, utilizing hardware performance counters
  - **FLOPS count** is obtained using:
    - SDE FLOPS calculations: accurate estimate, but very slow
    - HW performance counters using LIKWID: inaccurate, but very fast

Impact of NUMA and SNC modes

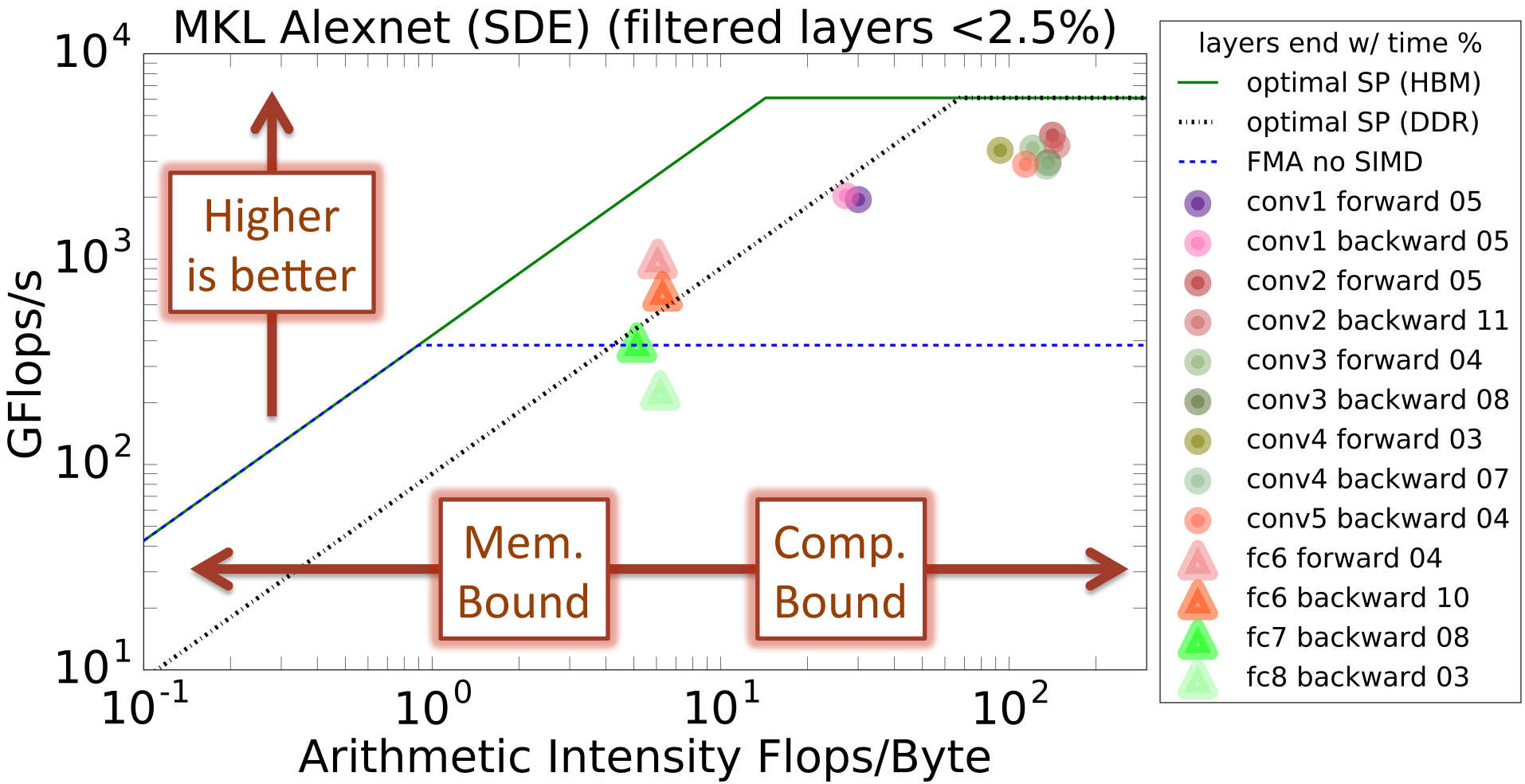
- **Obtained best performance using 1 thread/core**
  - Leaving 1 thread for the OS achieves 5% better performance
- **Same performance is obtained with:**
  - **Quad-flat** vs. **SNC4-flat** mode (tested in Bin-1 KNL 7250 @1.4GHz)
  - **Quad-flat** vs. **Quad-cache** mode (tested in Bin-3 KNL 7210 @1.3GHz)
- **Binding to HBM is ~60% faster than to DRAM**

Setup:

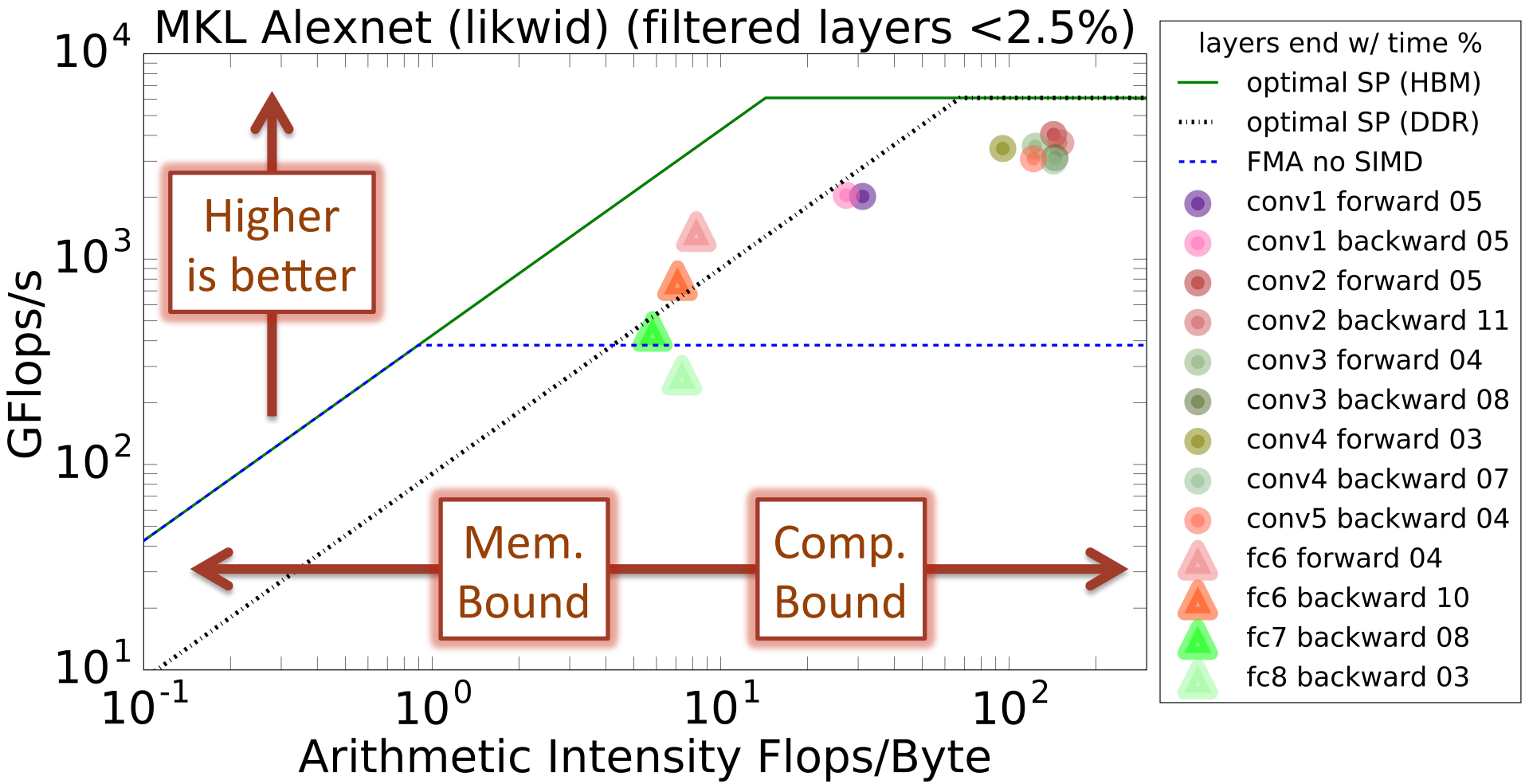
- **Intel Caffe**
- **MKL Alexnet DNN arch.**



Kernels Roofline breakdown – with SDE



Kernels Roofline breakdown – with LIKWID



Thank you