Extending the BLIS Analytical Model for GPUs

Elliot Binder, Claudia Kho, Doru Thom Popovici, Tze Meng Low

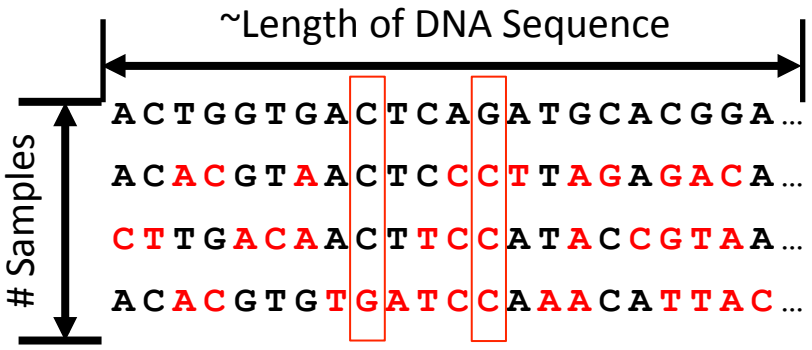
18 September 2018

BLIS Retreat

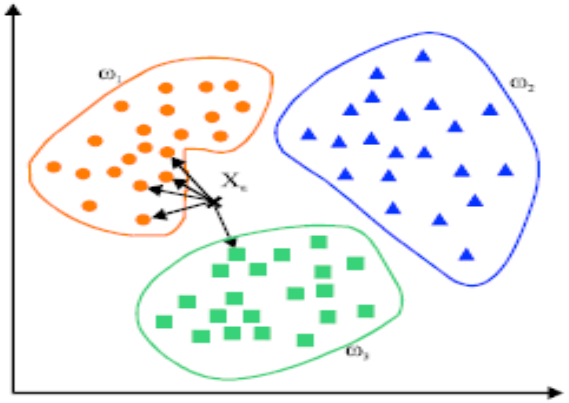
Electrical & Computer ENGINEERING

Many problems are "MMM"

Population Genomics

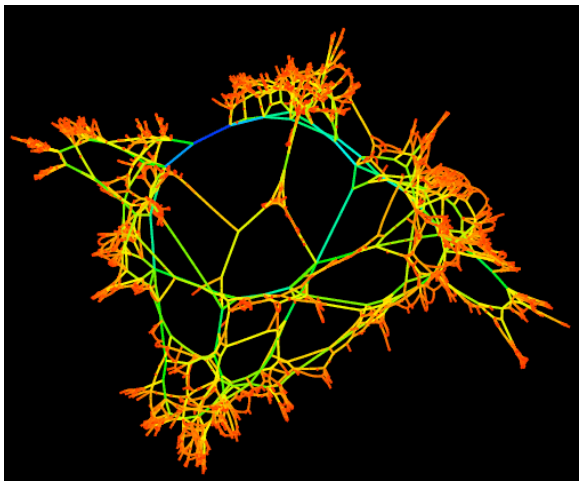


k-Nearest Neighbours



DNA Fingerprinting

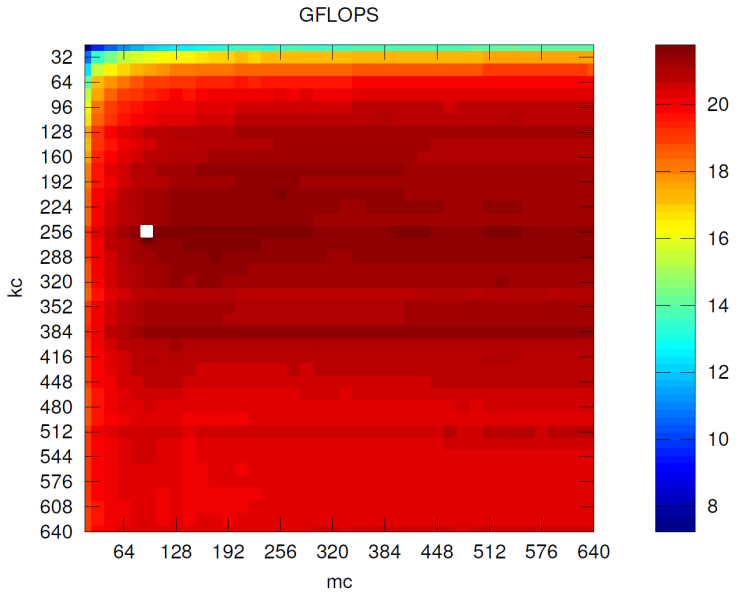
All-Pairs Shortest Path

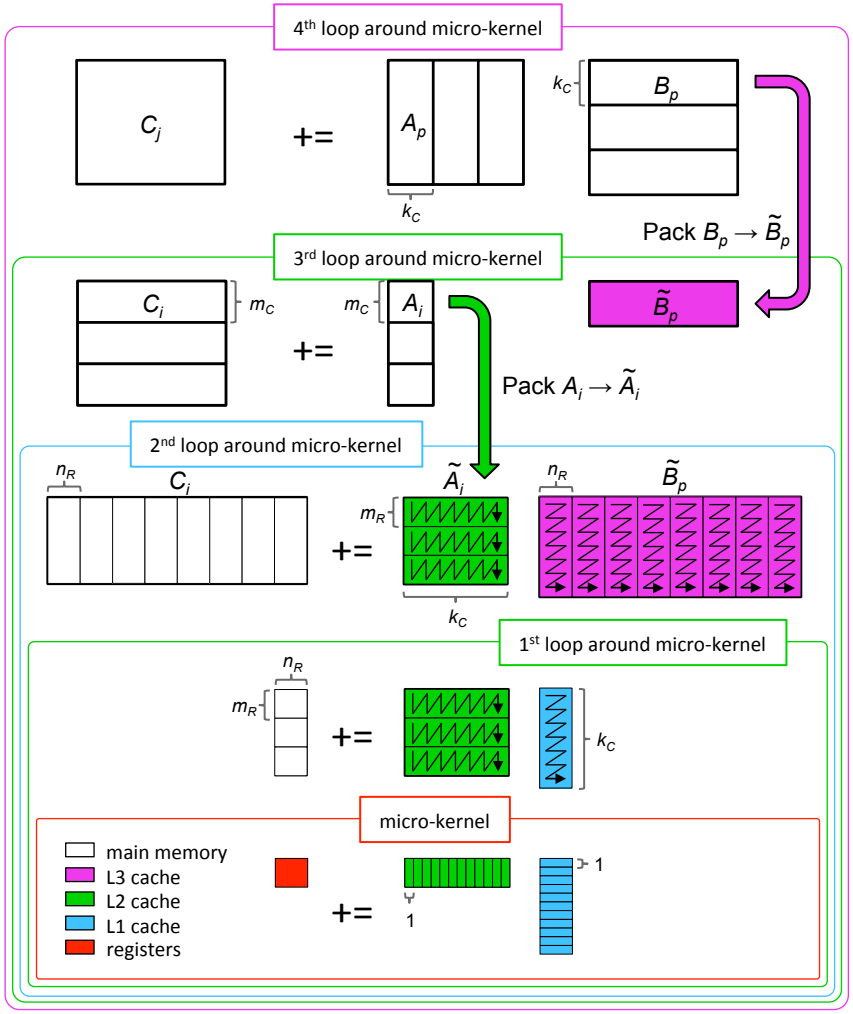


Leveraging BLIS

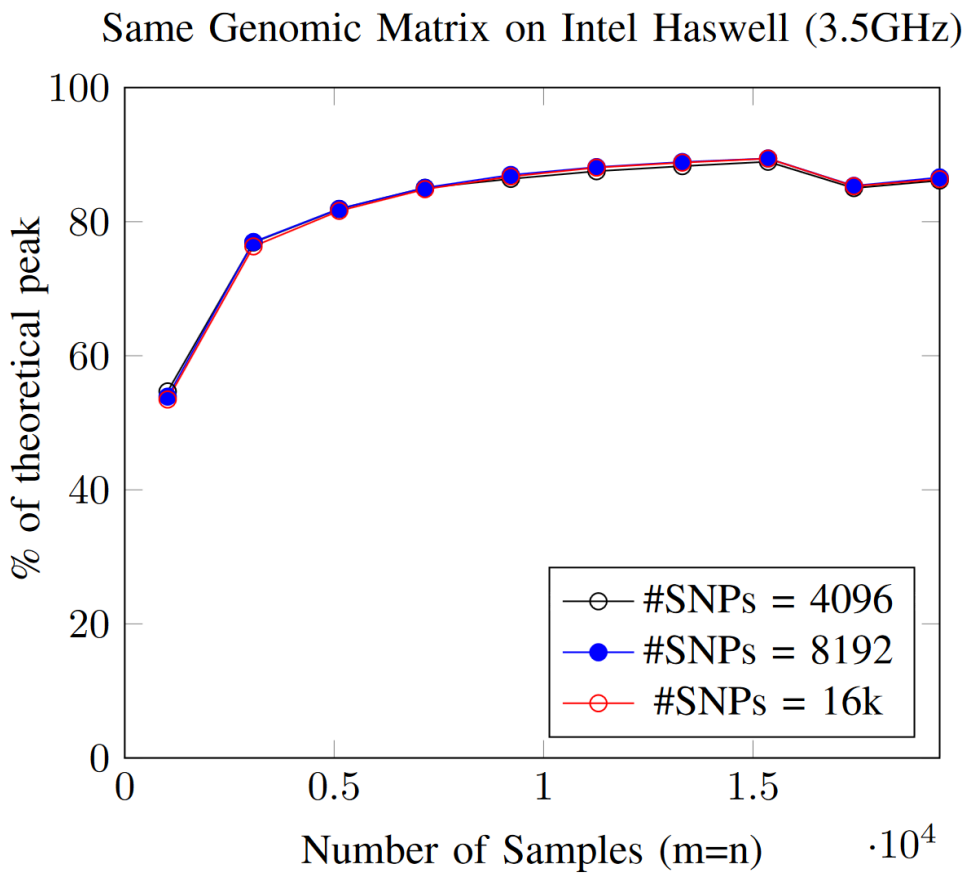
- Small microkernel
- 5 parameters

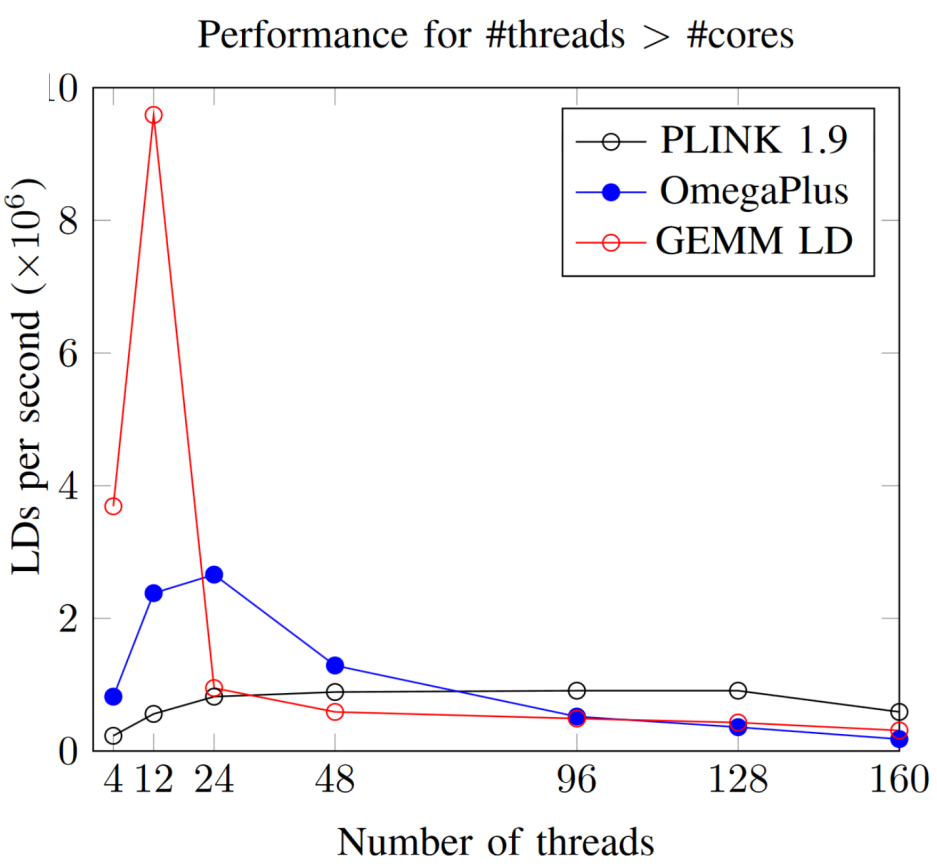
$$m_r \;\; n_r \;\; k_c \;\; m_c \;\; n_c$$





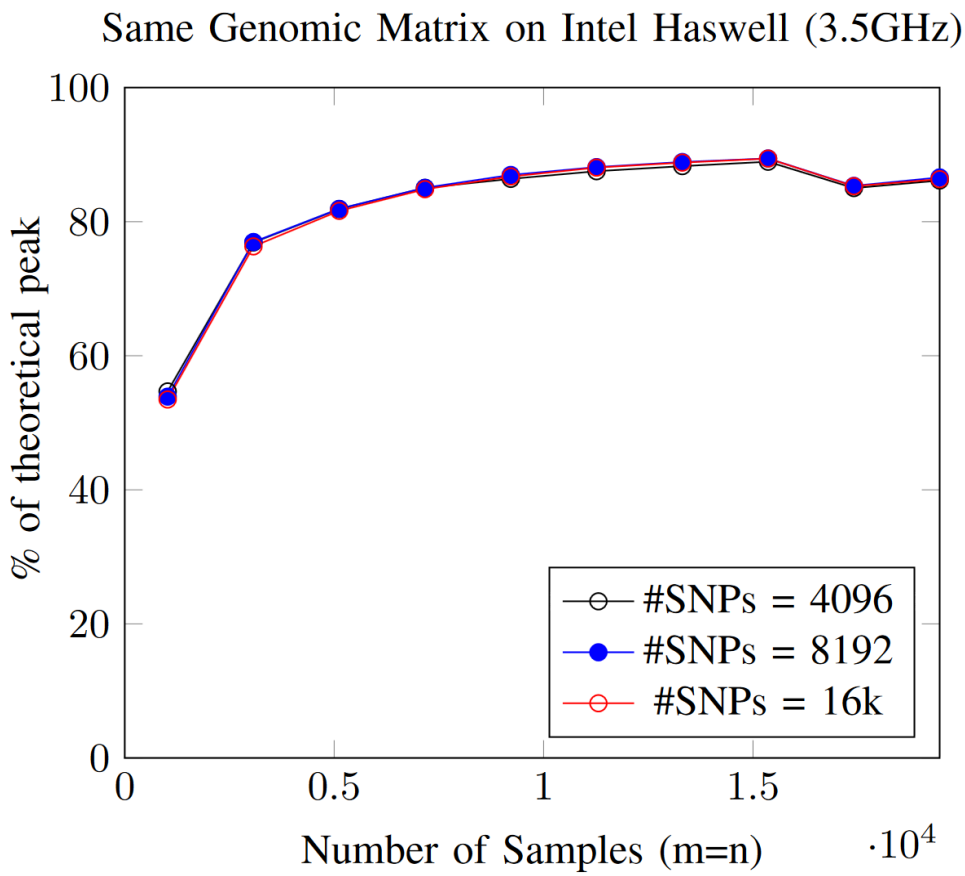
Population Genomics

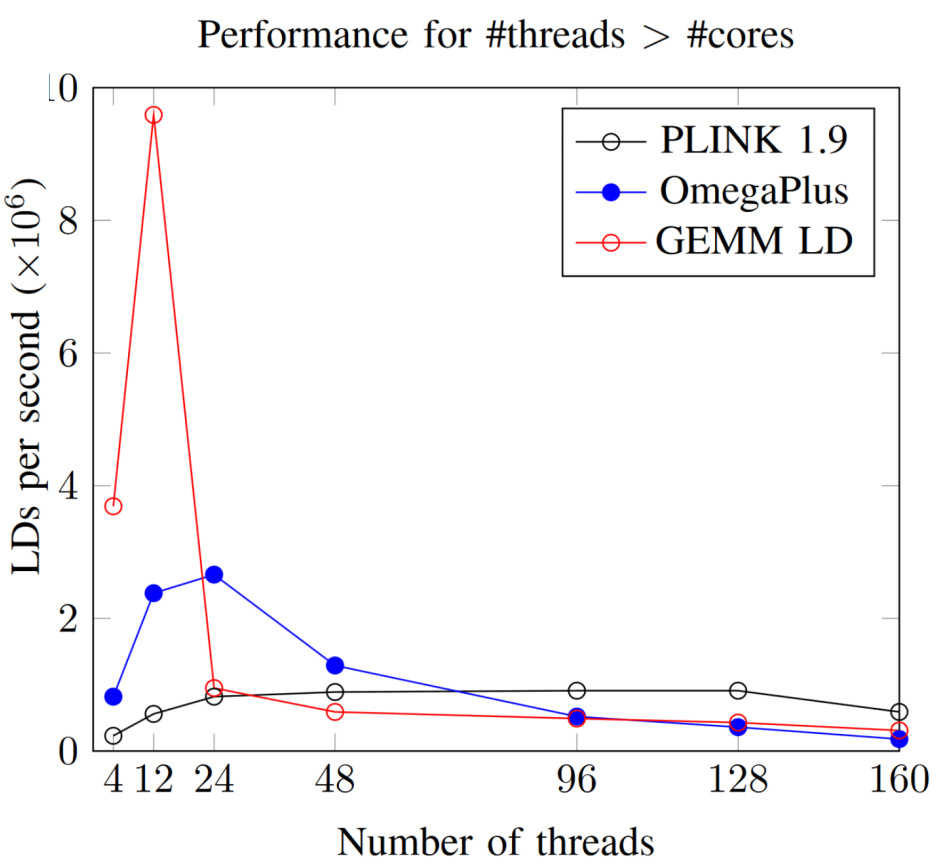




Nikolaos Alachiotis, Thom Popovici, Tze Meng Low, 2016. Efficient Computation of Linkage Disequilibria as Dense Linear Algebra Operations. HiCOMB 2016

Population Genomics





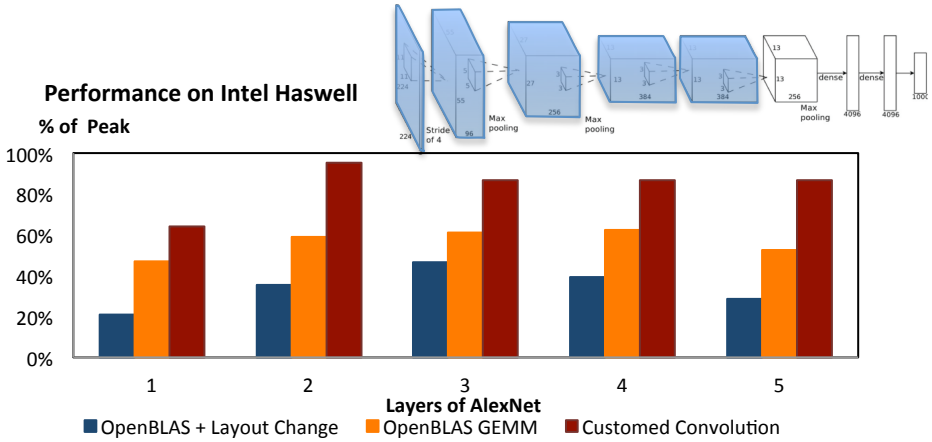
$$m_r n_r \geq N_{\text{Popcnt}} L_{\text{Popcnt}} N_{vec}$$

Tze Meng Low, Francisco D. Igual, Tyler M. Smith, and Enrique S. Quintana-Orti. 2016. Analytical Modeling Is Enough for High-Performance BLIS. *ACM Trans. Math. Softw.* 43, 2, Article 12

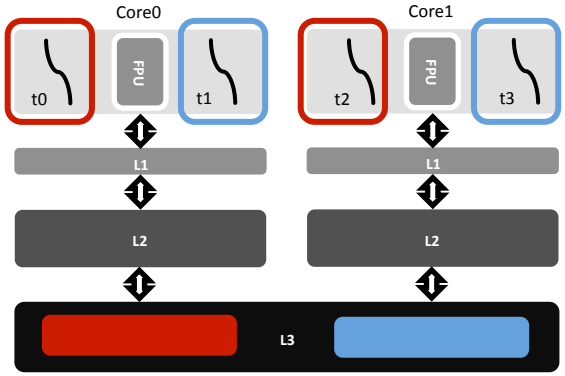
Nikolaos Alachiotis, Thom Popovici, Tze Meng Low, 2016. Efficient Computation of Linkage Disequilibria as Dense Linear Algebra Operations. HiCOMB 2016

Application of (Partial) Model

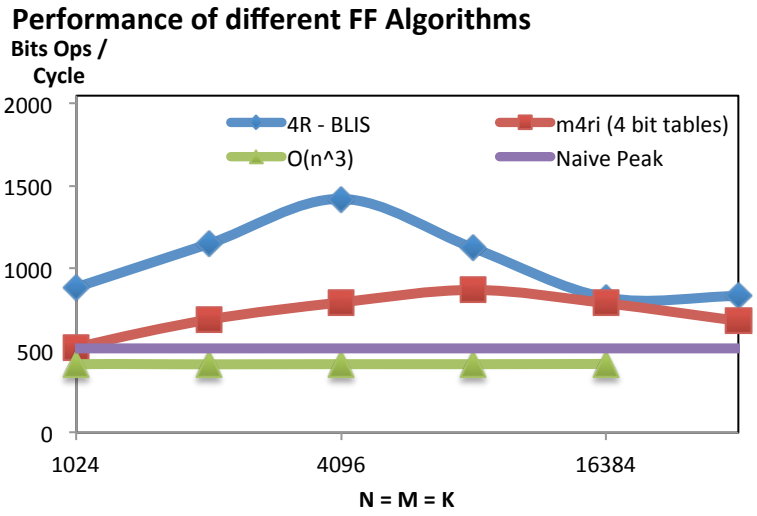
Convolution Neural Nets



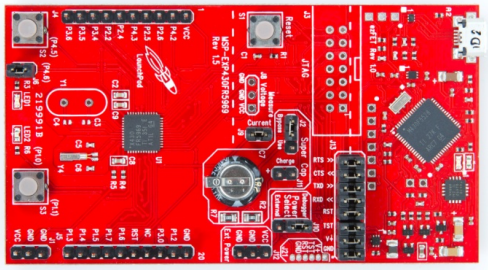
Large m-D FFTs



Finite Field Linear Algebra

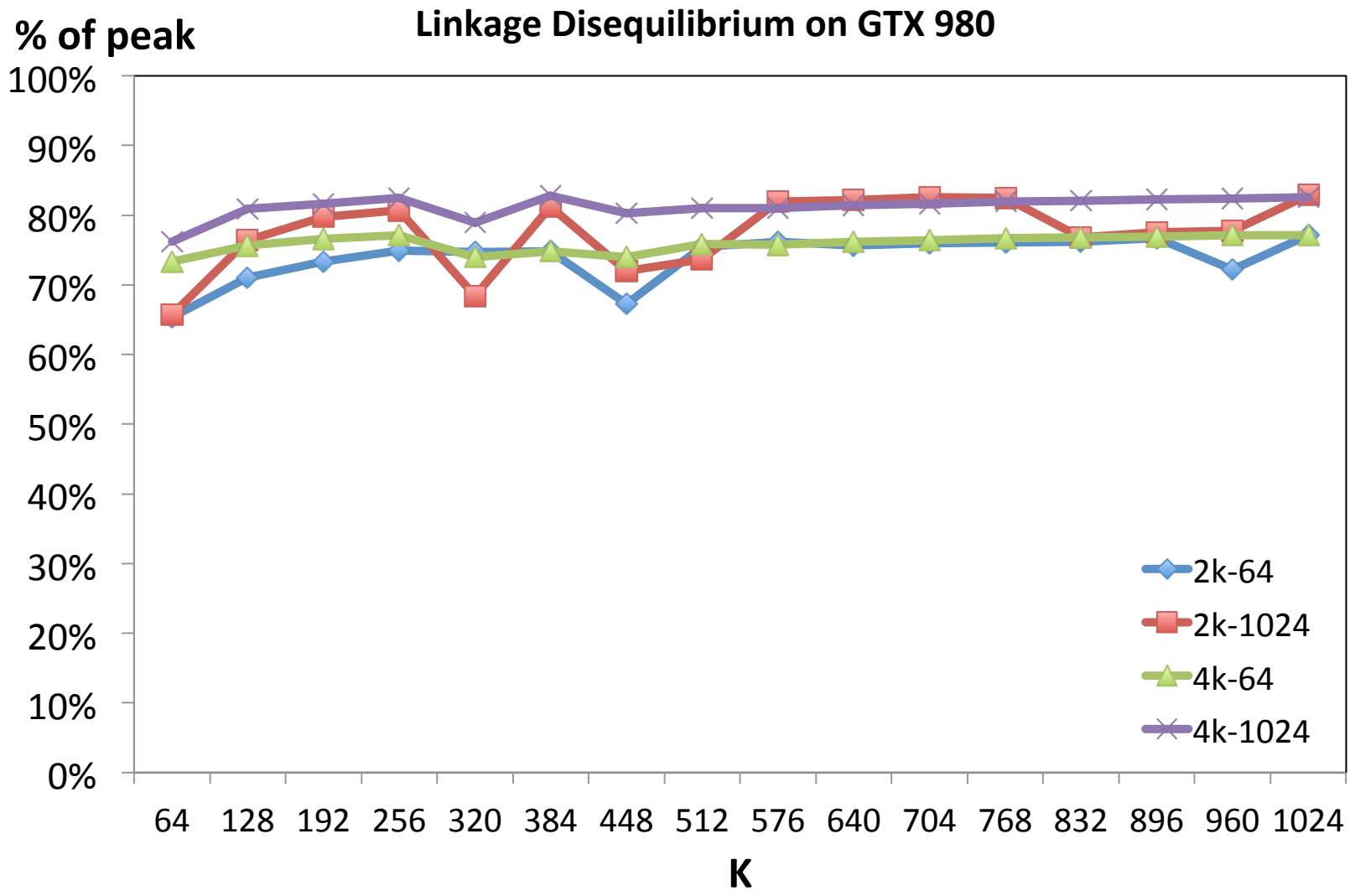


Microcontrollers



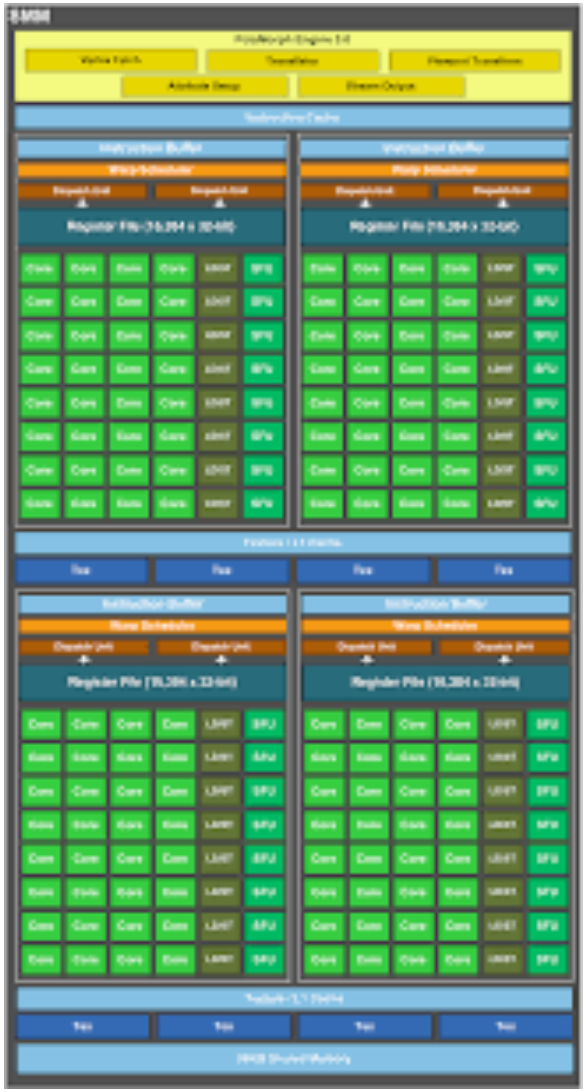
"Can we do it on a GPU?"

Our initial attempt



GTX 980 in a nutshell

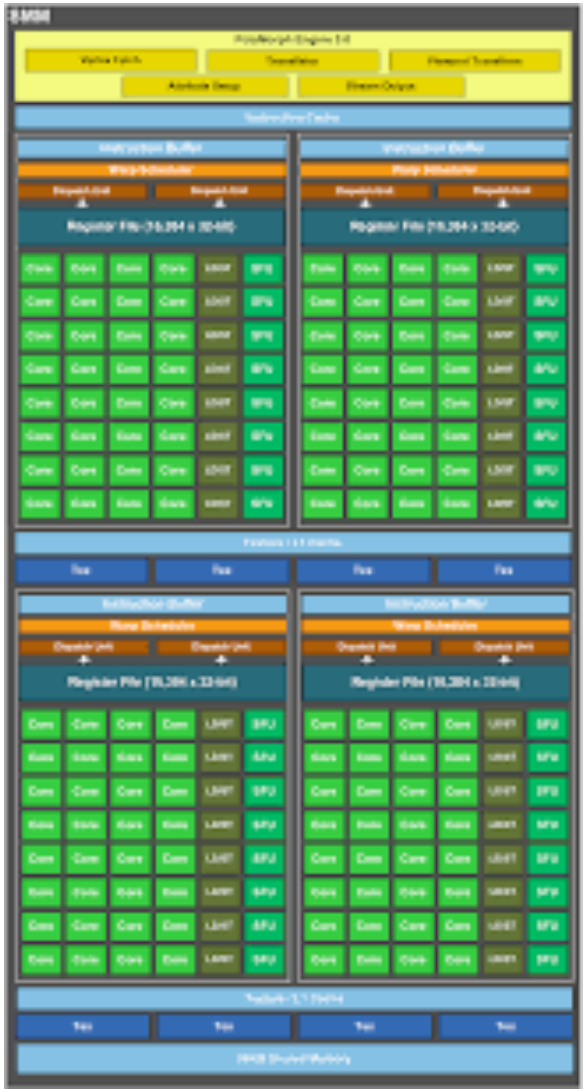
1 of 16 SMs



- 1 warp = 32 threads
- 4 clusters of 32 (SP) FMA cores
- Each cluster with 8 SFU cores (popcnt)
- 64k registers per SM (255/thread)
- 48K/96K shared memory

GTX 980 in a nutshell

1 of 16 SMs



- 1 warp = 32 threads
- 4 clusters of 32 (SP) FMA cores
- Each cluster with 8 SFU cores (popcnt)
- 64k registers per SM (255/thread)
- 48K/96K shared memory
- Latency of FMA ≈ 8 cycles
- Latency of Popcnt ≈ 12-13 cycles
- Popcnt seems to be pipelined

Applying the model

- Minimum size of kernel

$$m_r n_r \geq N_{\mathrm{Popcnt}} L_{\mathrm{Popcnt}} N_{vec}$$

256 4 clusters 8 cycles 8 threads

- Maximum size of kernel

$$\frac{64k}{256} = 256$$

Applying the model

- Minimum size of kernel

$$m_r n_r \geq N_{\text{Popcnt}} L_{\text{Popcnt}} N_{vec}$$

256 4 clusters 8 cycles 8 threads

- Maximum size of kernel

$$\frac{64k}{256} = \boxed{256}$$

>255 registers/thread

Applying the model

- Minimum size of kernel

$$m_r n_r \geq N_{\text{Popcnt}} L_{\text{Popcnt}} N_{vec}$$

256 4 clusters 8 cycles 8 threads

- Maximum size of kernel

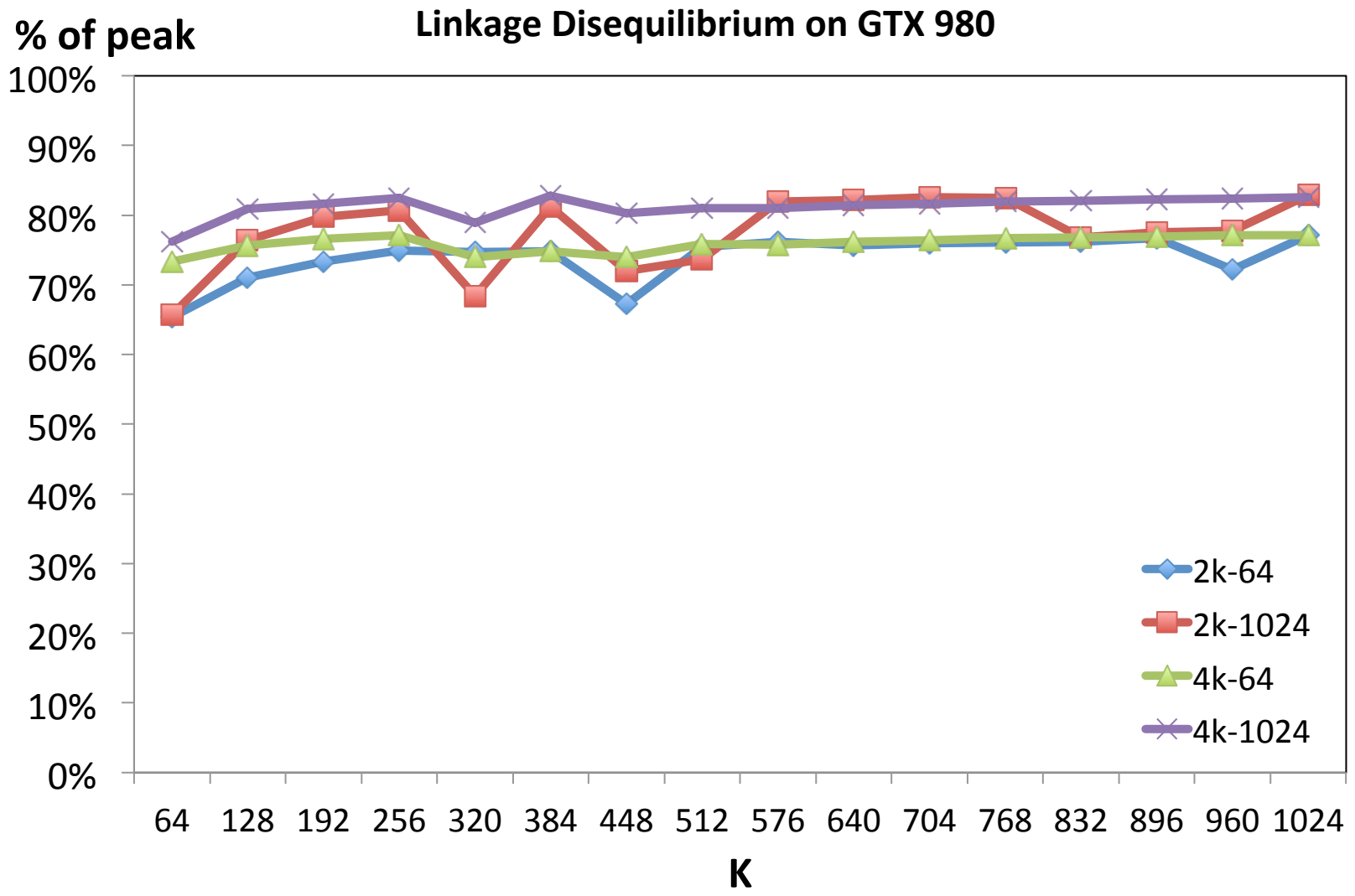
$$\frac{64k}{256} = 256$$

1024

↑ Threads, ↓ Registers

64

Our initial attempt



With Shared Memory