BLIS Performs

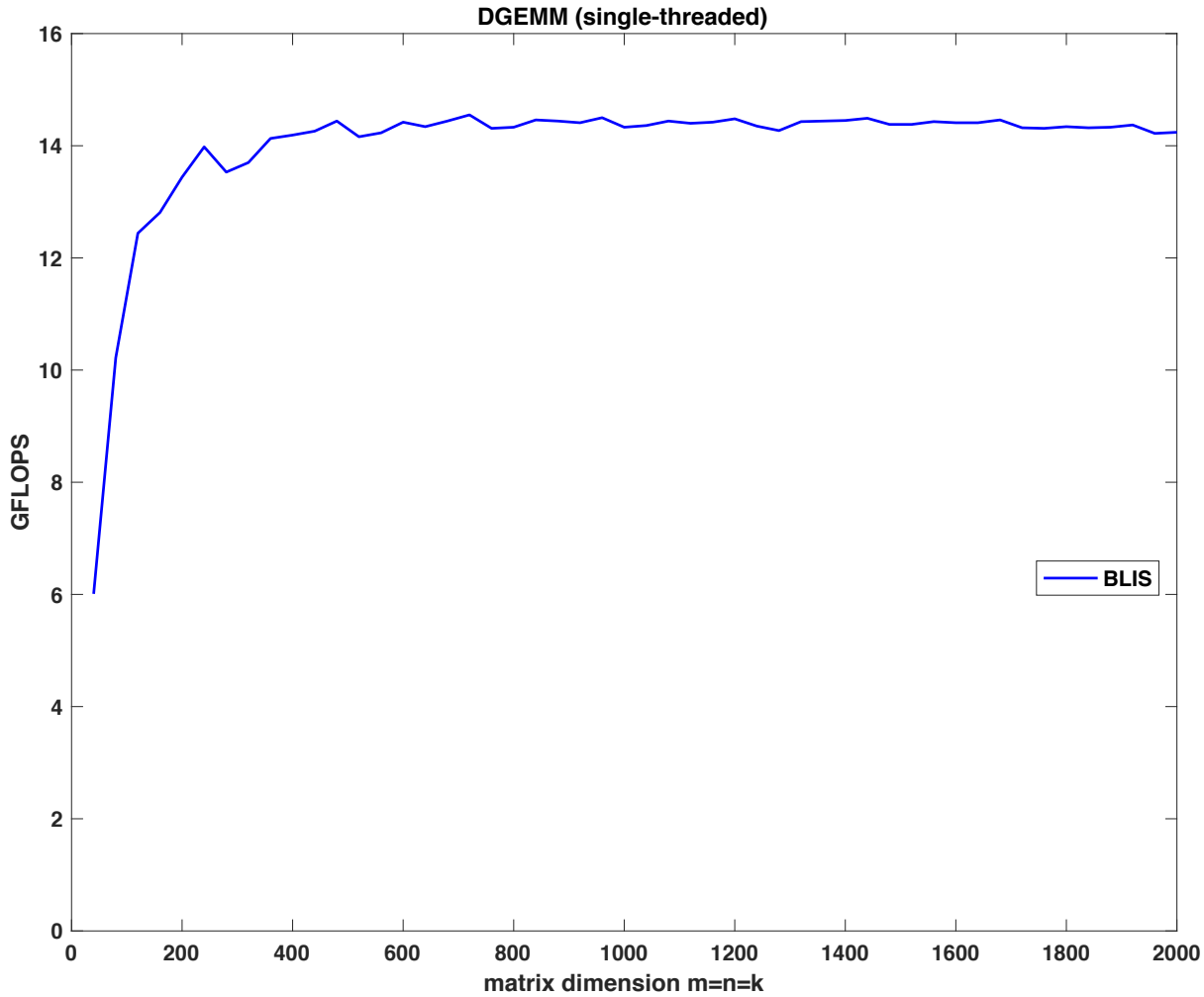
Devangi N. Parikh
Science of High Performance Computing
The University of Texas at Austin

ThunderX2

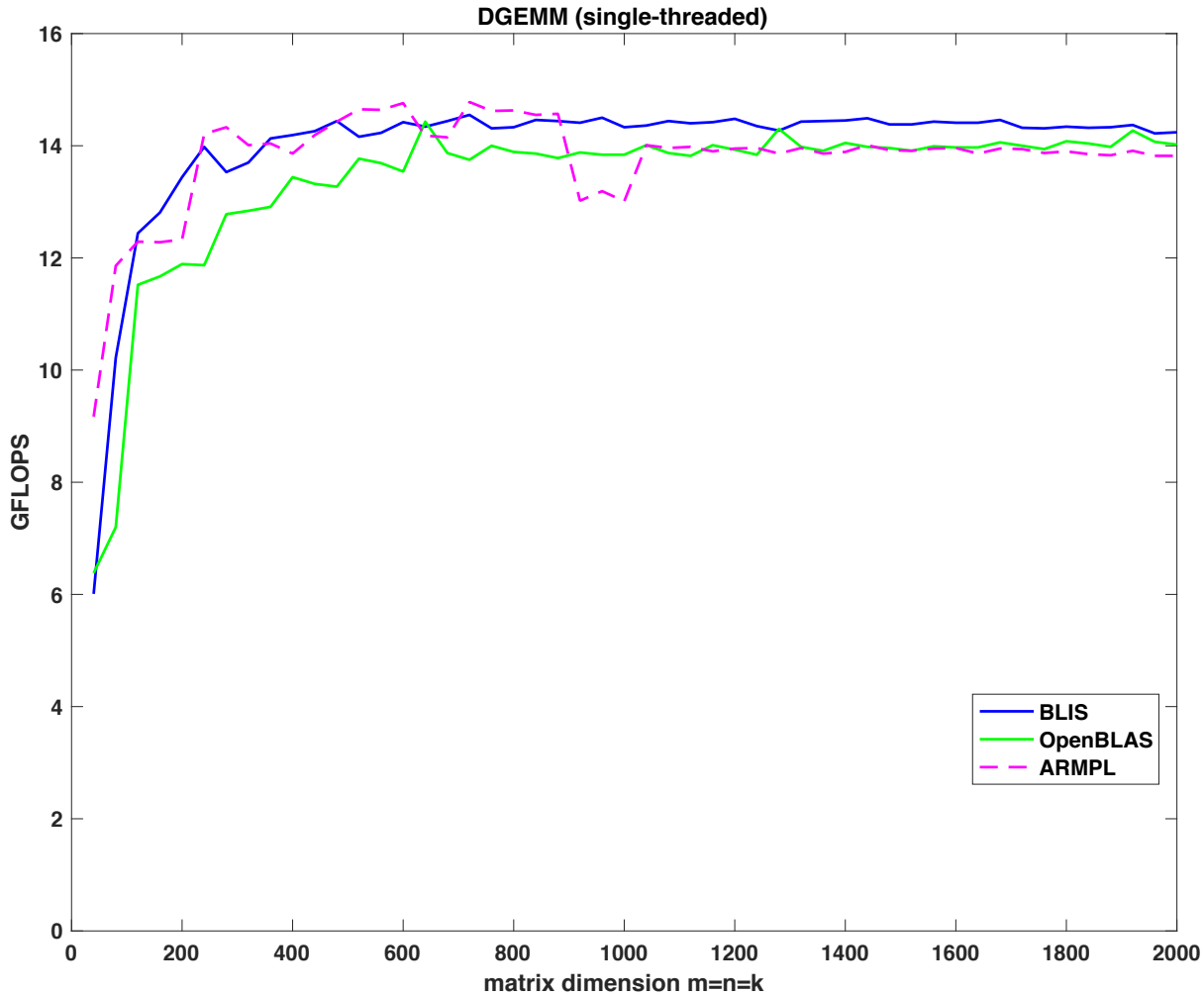
Architecture	arm v8.1
Base frequency	2.0 GHz
# sockets/node	2
# cores/socket	28

armv8a kernels in BLIS were written by Fransisco D. Igual for cortexa57 architectures.

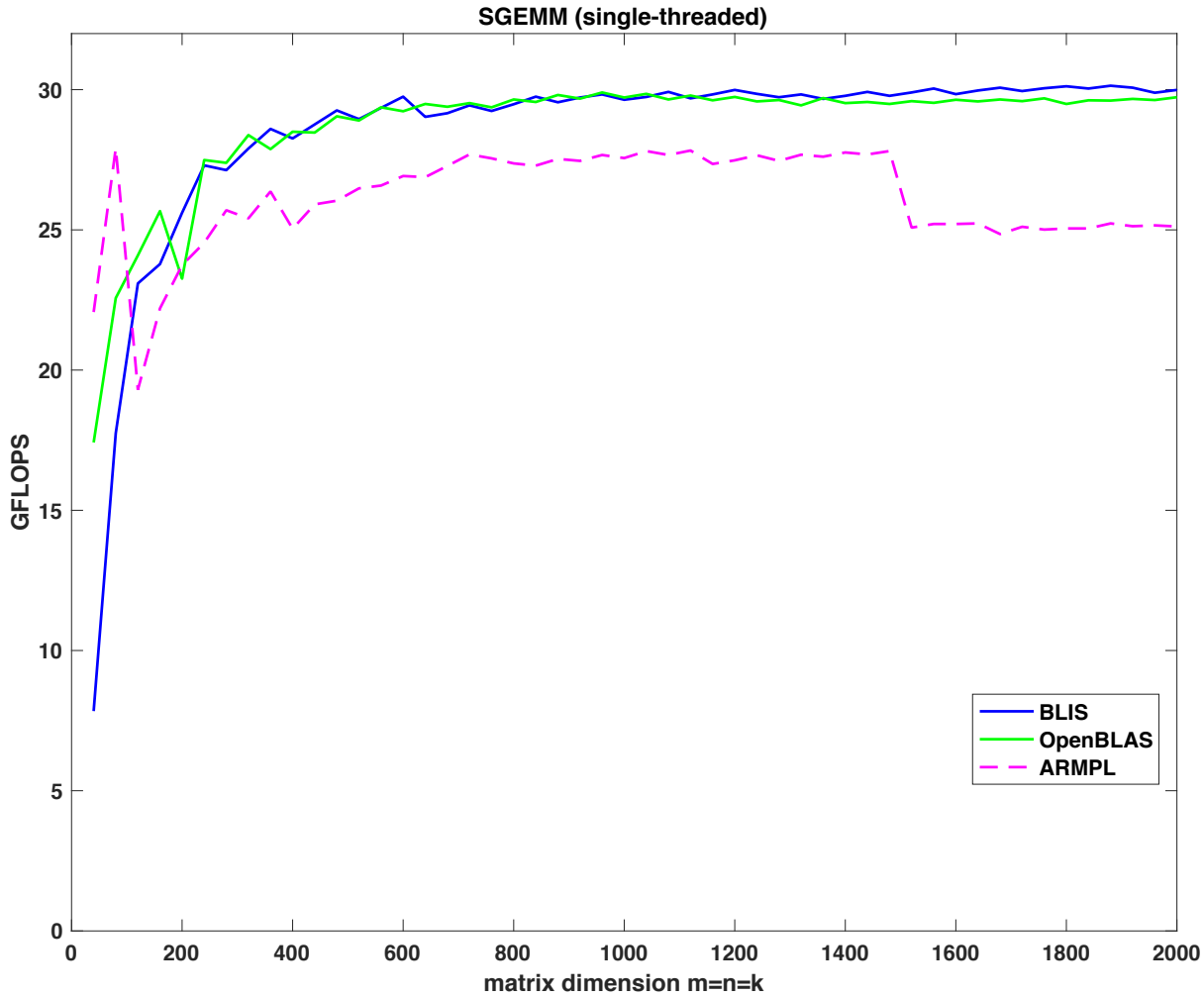
DGEMM (armv8a)



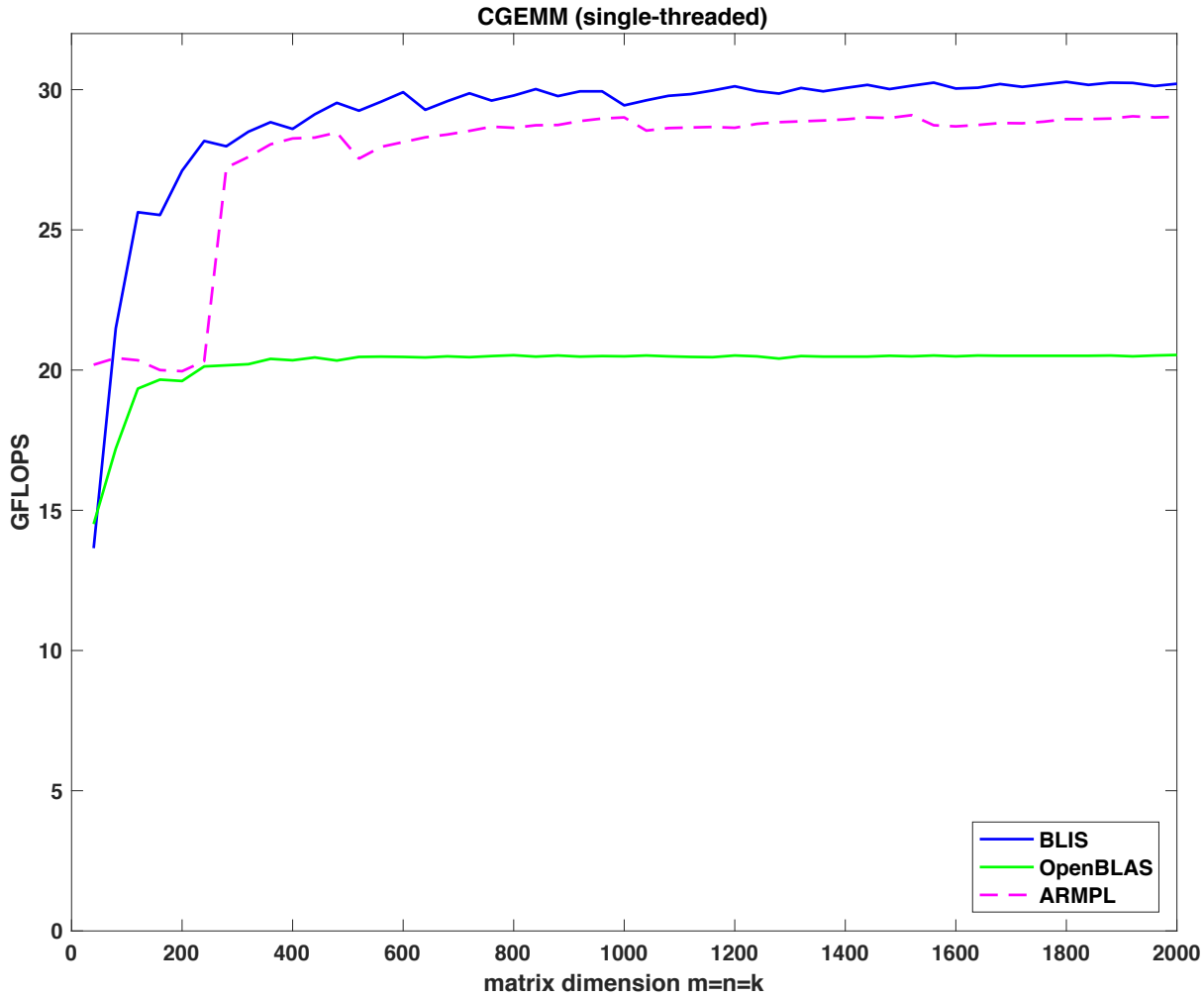
DGEMM – Other Libraries



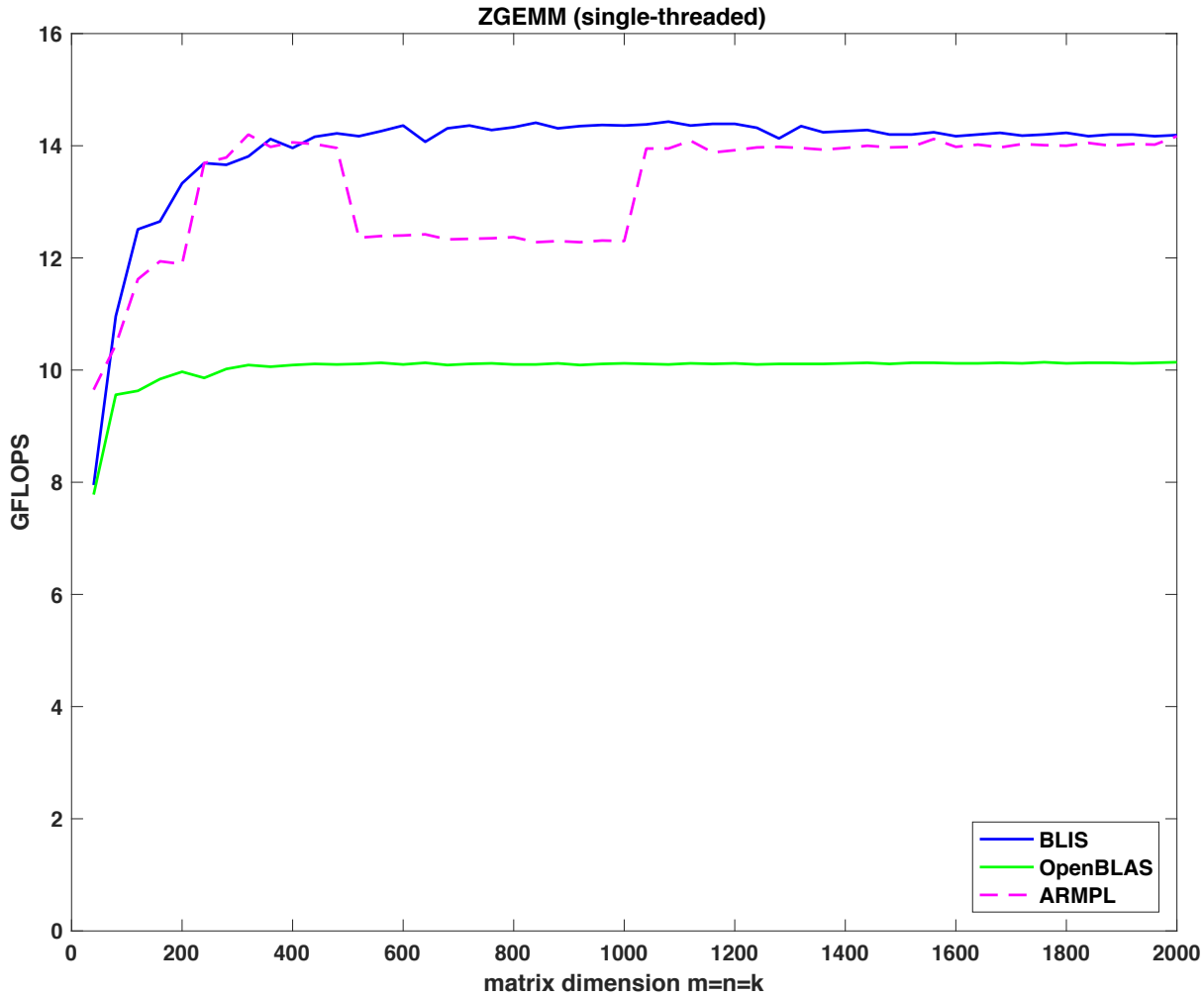
GEMM – Other Datatypes



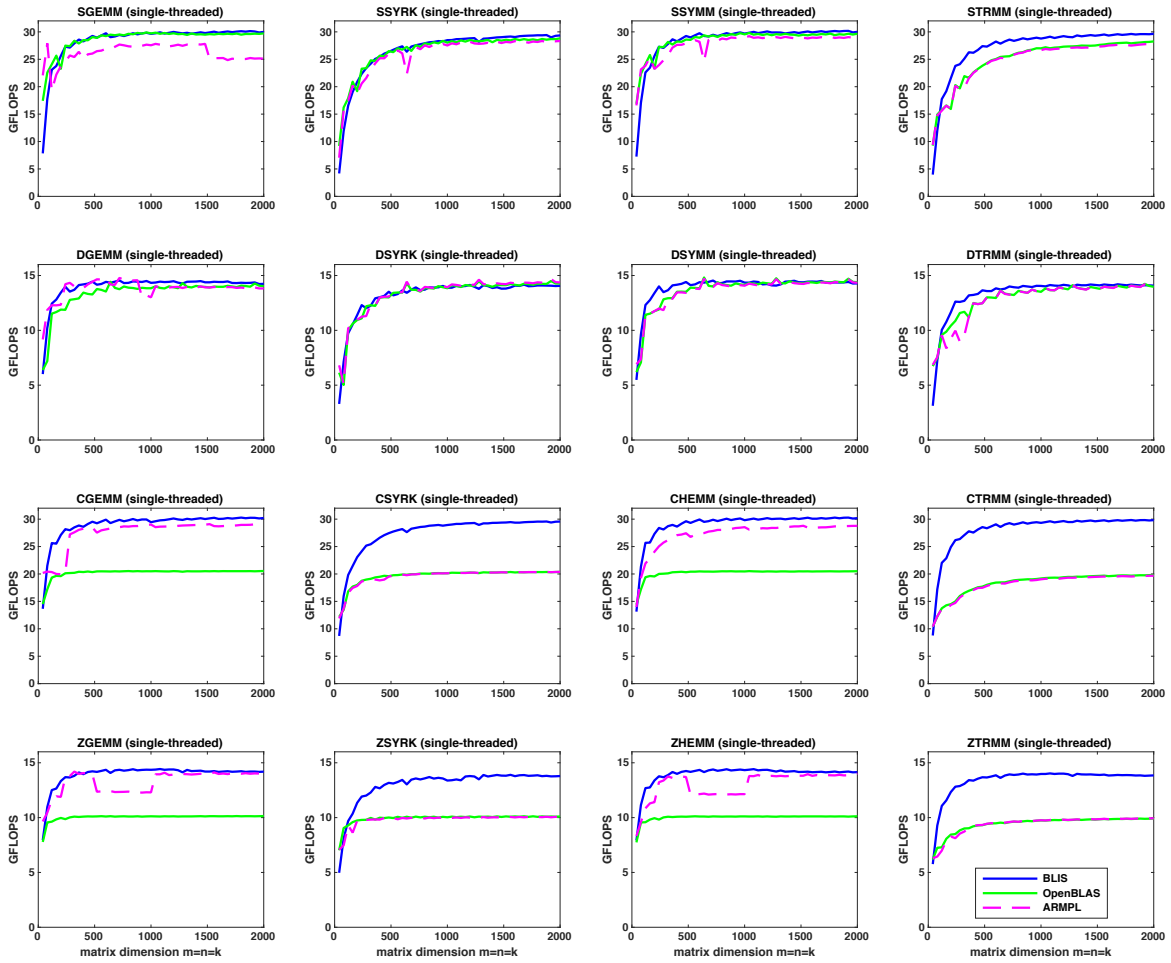
GEMM – Other Datatypes



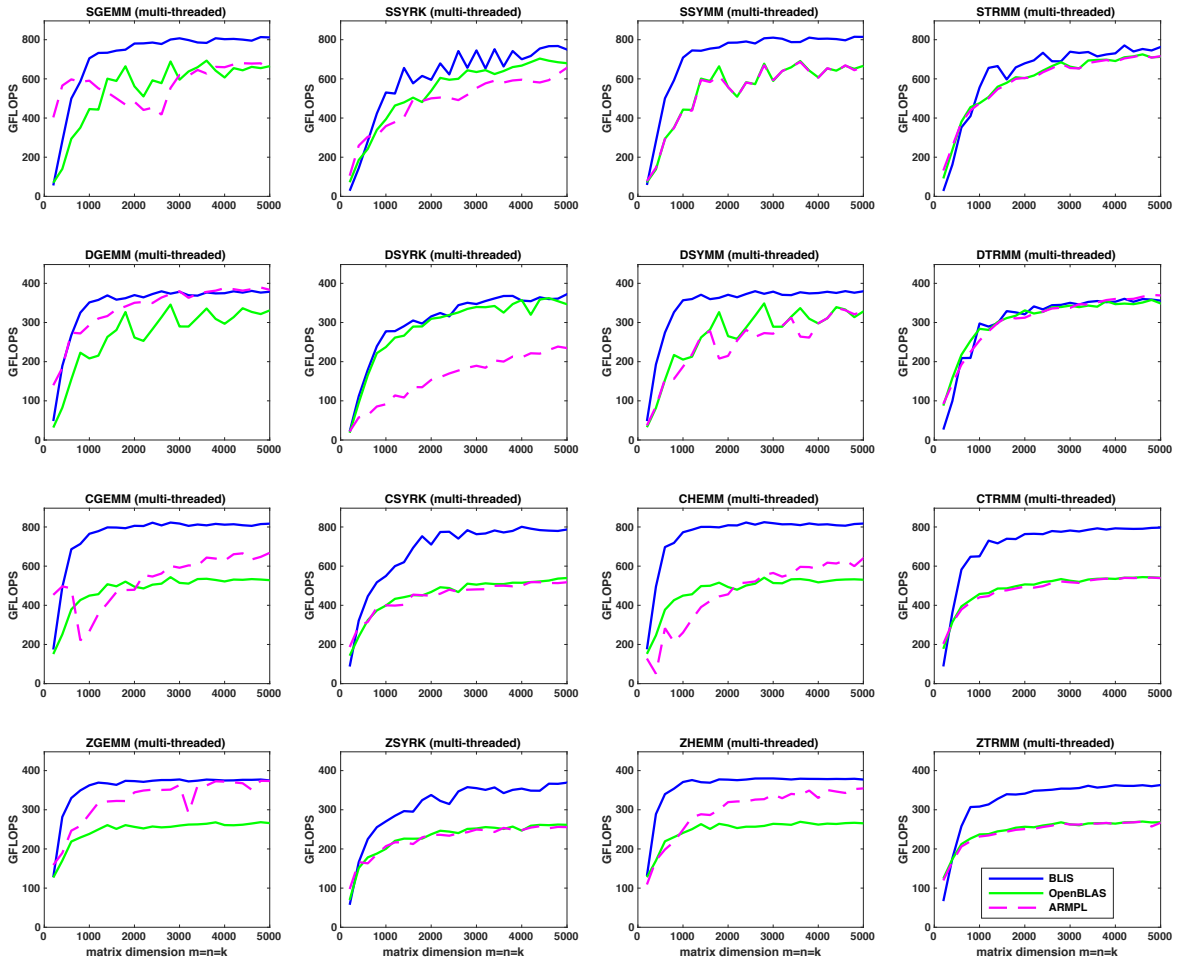
GEMM – Other Datatypes



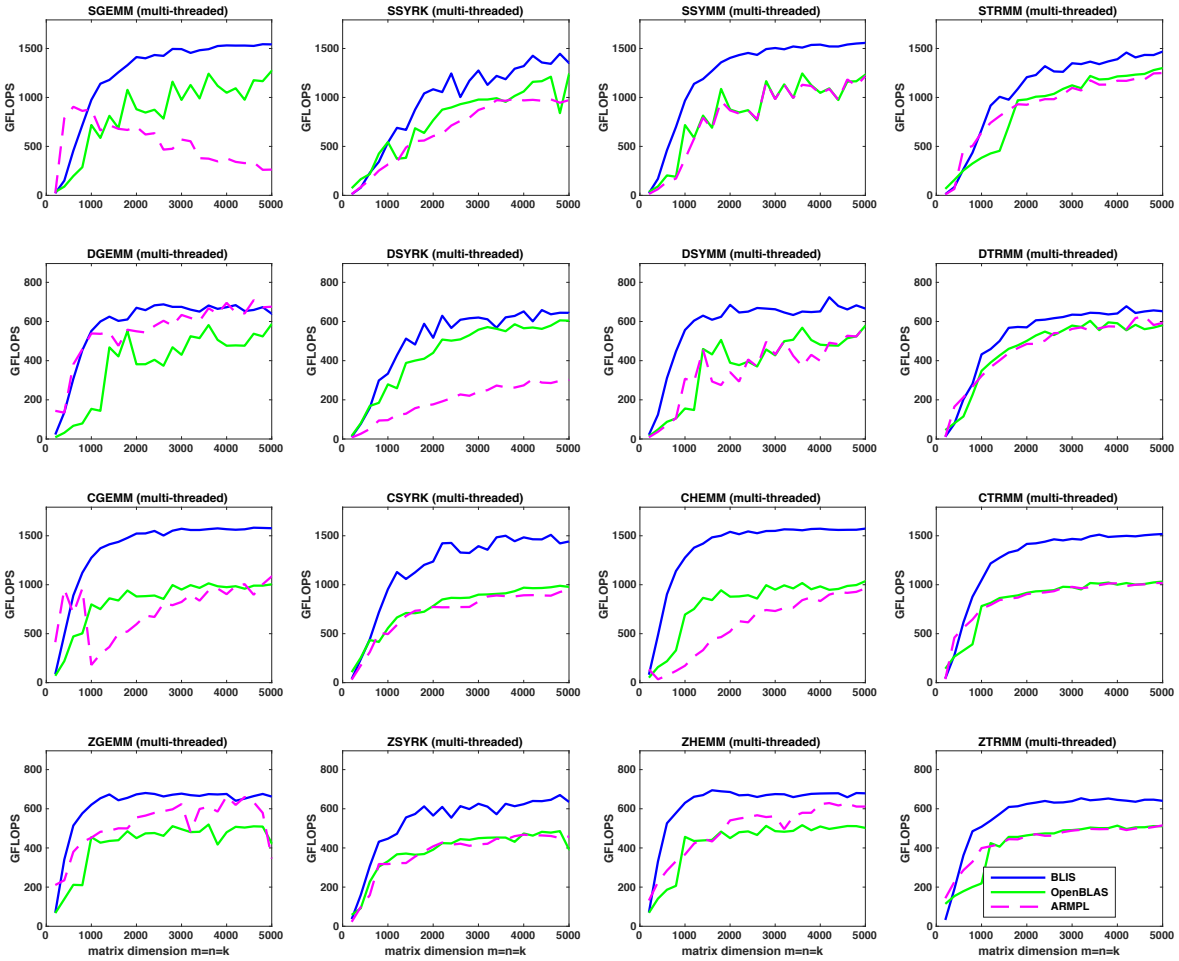
Level 3


SGEMM (single-threaded)
SSYRK (single-threaded)
SSYMM (single-threaded)
STRMM (single-threaded)
DGEMM (single-threaded)
DSYRK (single-threaded)
DSYMM (single-threaded)
DTRMM (single-threaded)
CGEMM (single-threaded)
CSYRK (single-threaded)
CHEMM (single-threaded)
CTRMM (single-threaded)
ZGEMM (single-threaded)
ZSYRK (single-threaded)
ZHEMM (single-threaded)
ZTRMM (single-threaded)
GFLOPS
matrix dimension m=n=k
BLIS
OpenBLAS
ARMPL

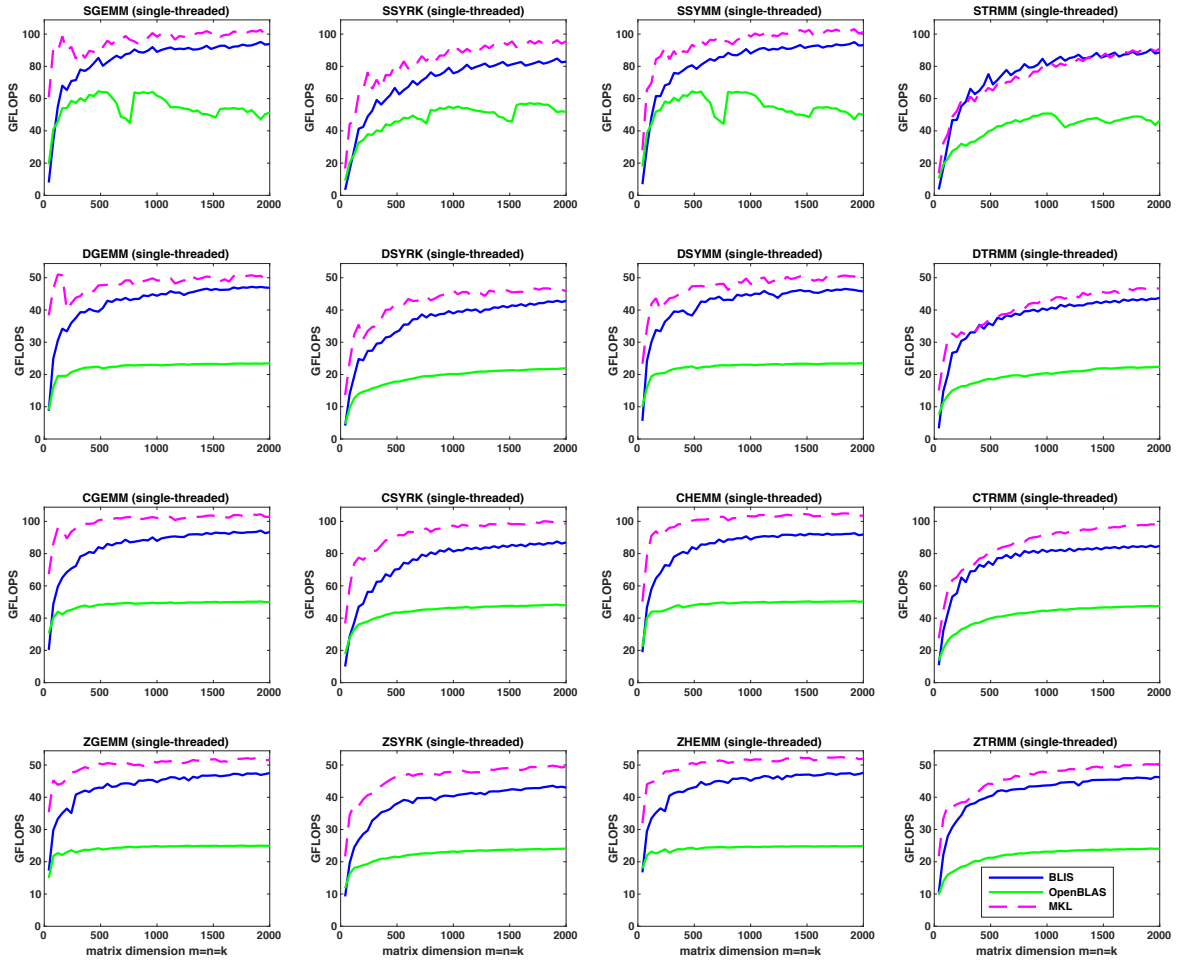
Multi-threaded BLIS (28 cores)



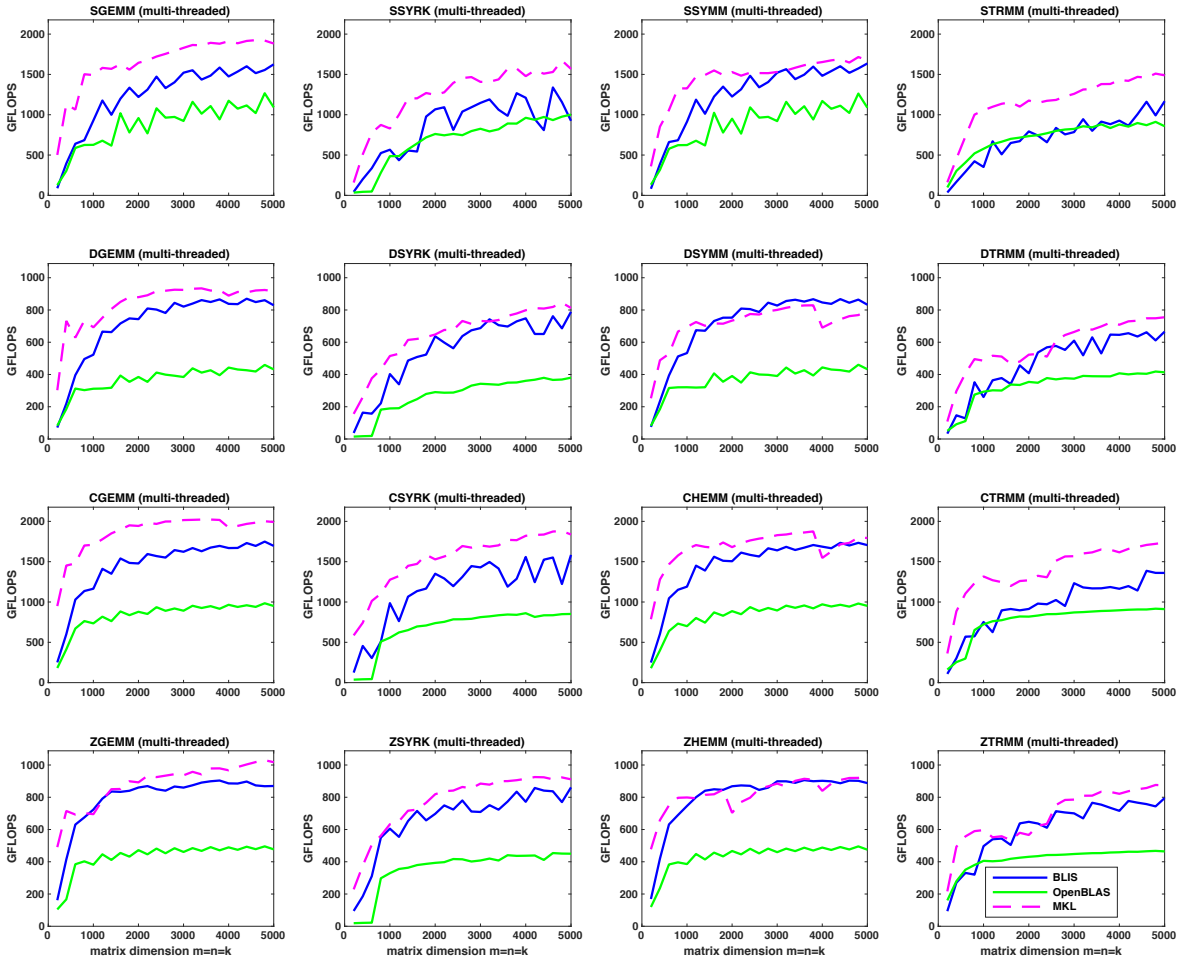
Multi-threaded BLIS (56 cores)



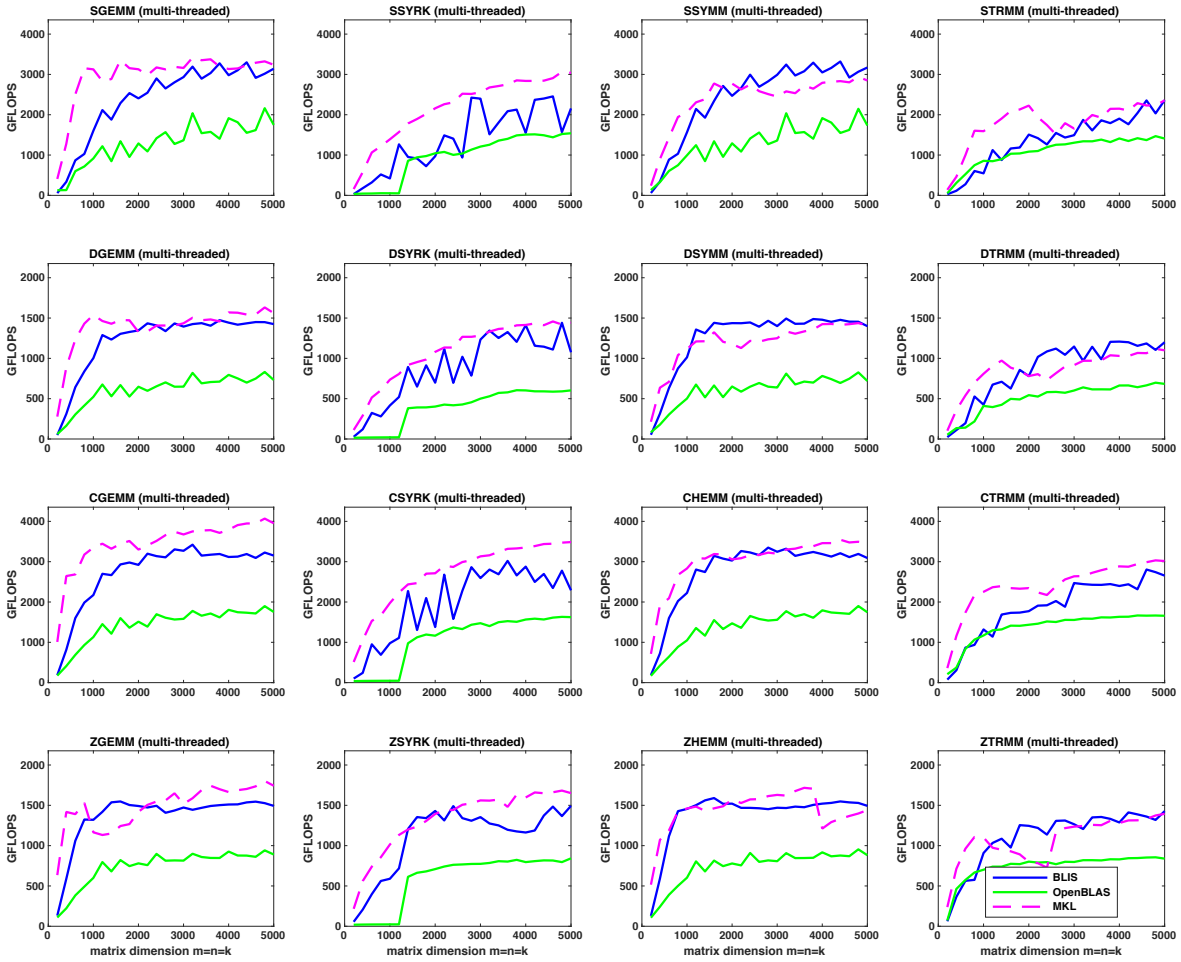
Other Architectures
SkylakeX (single core)



Other Architectures
SkylakeX (20 cores)



Other Architectures
SkylakeX (40 cores)



Conclusion

BLIS instantiates high-performance implementations across virtually all level-3 operations, parameter cases, and datatypes with just two microkernels.

- sgemm
- dgemm

Thank you!

- Thanks to Fransisco D. Igual for running experiments on a frequency stable SkylakeX.
- Thanks to Cavium and ARM for access to ThunderX2.

(More plots on poster)

Details

- ThunderX2
  - 2.0 GHz
  - 2 sockets
  - 28 cores/socket
  - Hardware threads turned off
  - OpenBLAS (commit 52d3f7a), built for target THUNDERX2T99
  - ARMPL: armpl-18.4.0_ThunderX2CN99
  - 1 socket: JC_NT = 2; IC_NT = 14
  - 2 socket: JC_NT = 4; IC_NT = 14
- SkylakeX
  - 1.7 GHz
  - 2 sockets
  - 20 cores/socket
  - Frequency throttling turned off
  - OpenBLAS (commit 4dd70d9)
  - MKL 18
  - 1 socket: JC_NT = 1; IC_NT = 20
  - 2 socket: JC_NT = 2; IC_NT = 20
- Haswell
  - 3.2 GHz/3.0 GHz
  - 2 sockets
  - 12 cores/socket
  - MKL 11.3
  - 1 socket: JC_NT = 2; IC_NT = 6
  - 2 socket: JC_NT = 4; IC_NT = 6

- For single-threaded experiments, each graphs plots GFLOPS as a function of problem size--varied from 40 to 2000 in increments of 40--where all matrix operands' dimensions (m, n, k) are bound to the problem size. In other words, all matrices are square.
- For multi-threaded experiments, each graphs plots GFLOPS as a function of problem size--varied from 200 to 5000 in increments of 200–where all matrix operands' dimensions (m, n, k) are bound to the problem size. In other words, all matrices are square.
- The y-axis is scaled so that the top of the graphs correspond to theoretical peak for the clock rate.
- All results are performed on random data, with each point of each graph representing the best (shortest runtime) of three trials.
- I've omitted legends from all graphs except one in the bottom-right, only to minimize clutter.
- BLIS uses the 1M-induced methods for C,Z operations (ThunderX2, and SkylakeX)