



SAGE: Self-Tuning Approximation for Graphics Engines

Mehrzad Samadi¹, Janghaeng Lee¹, D. Anoushe Jamshidi¹, Amir Hormati², and Scott Mahlke¹

University of Michigan¹, Google Inc.²



cccq.eecs.umich.edu

Approximation is Acceptable in Many Domains

- Approximation
 - Machine Learning
 - Image Processing
 - Video processing
 - ...
- Less work
 - Higher performance
 - Lower power consumption

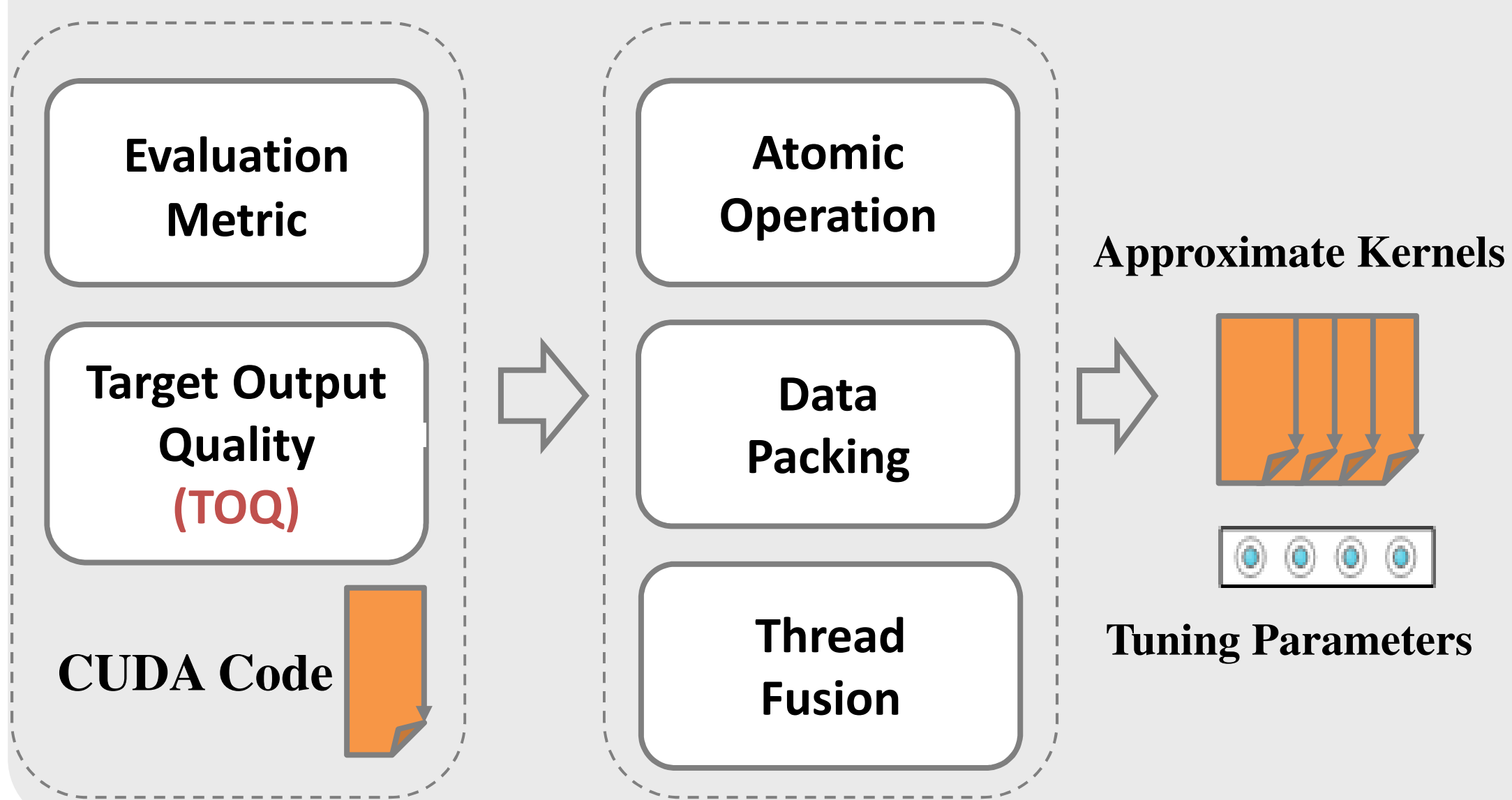


Improving Performance While Quality is Acceptable

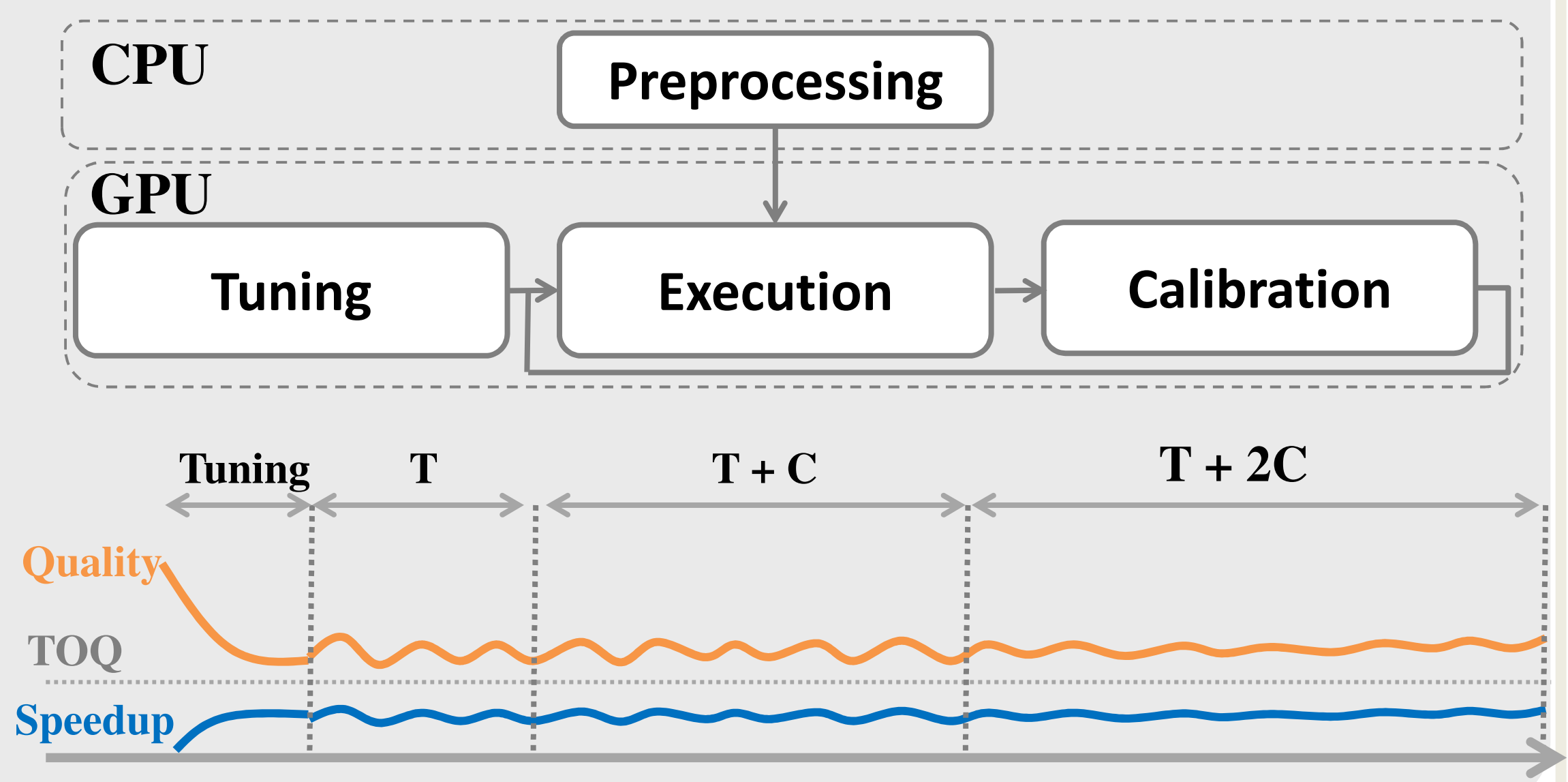
- Simplify or skip processing on the input data
 - Computationally expensive for GPU
 - Have the lowest impact on the output quality
- SAGE
 - Write the program once
 - Automatic approximation
 - Dynamic self-tuning



SAGE Generates Multiple Approximate Kernels

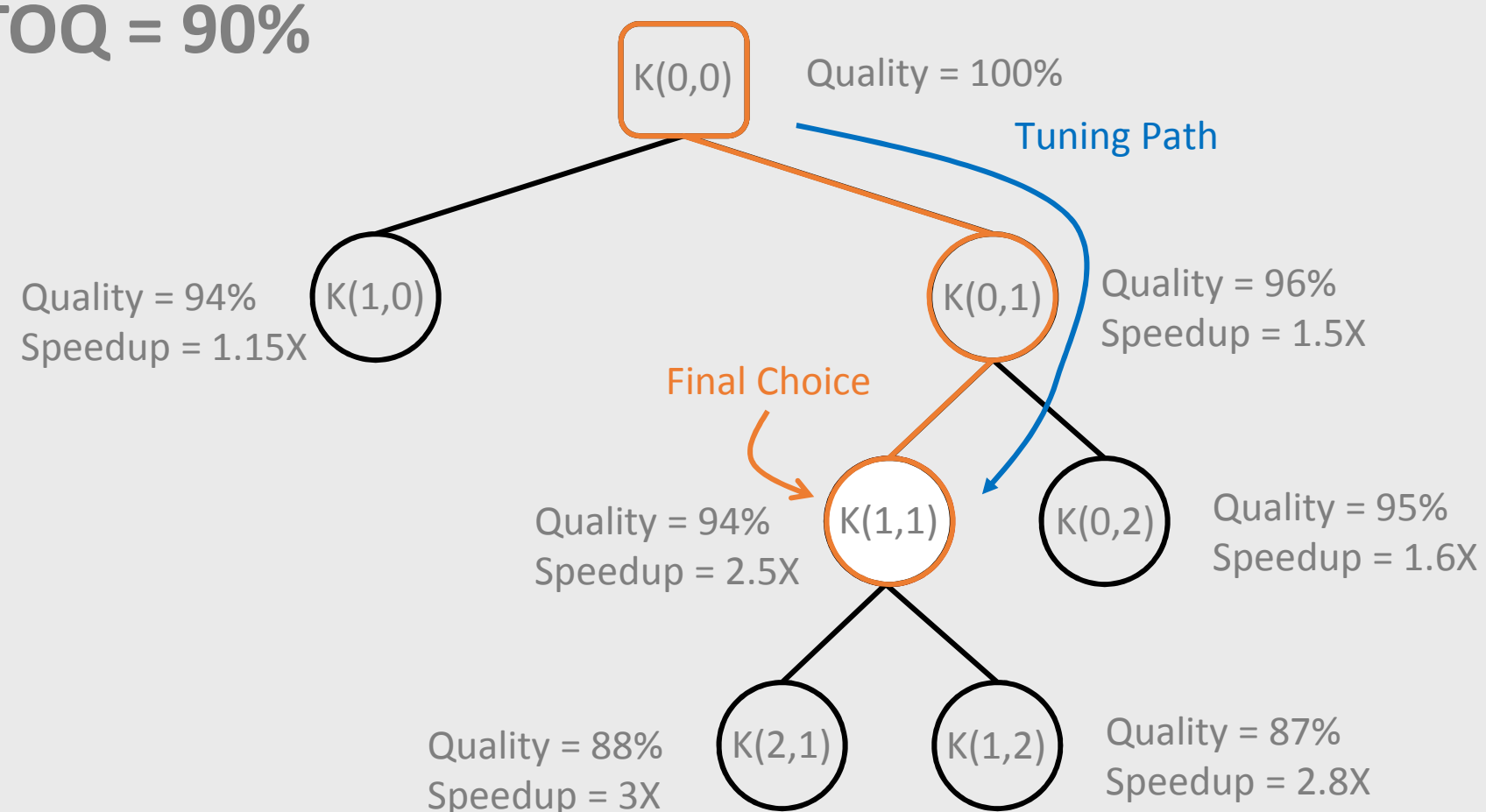


SAGE Monitors the Output Quality During Runtime

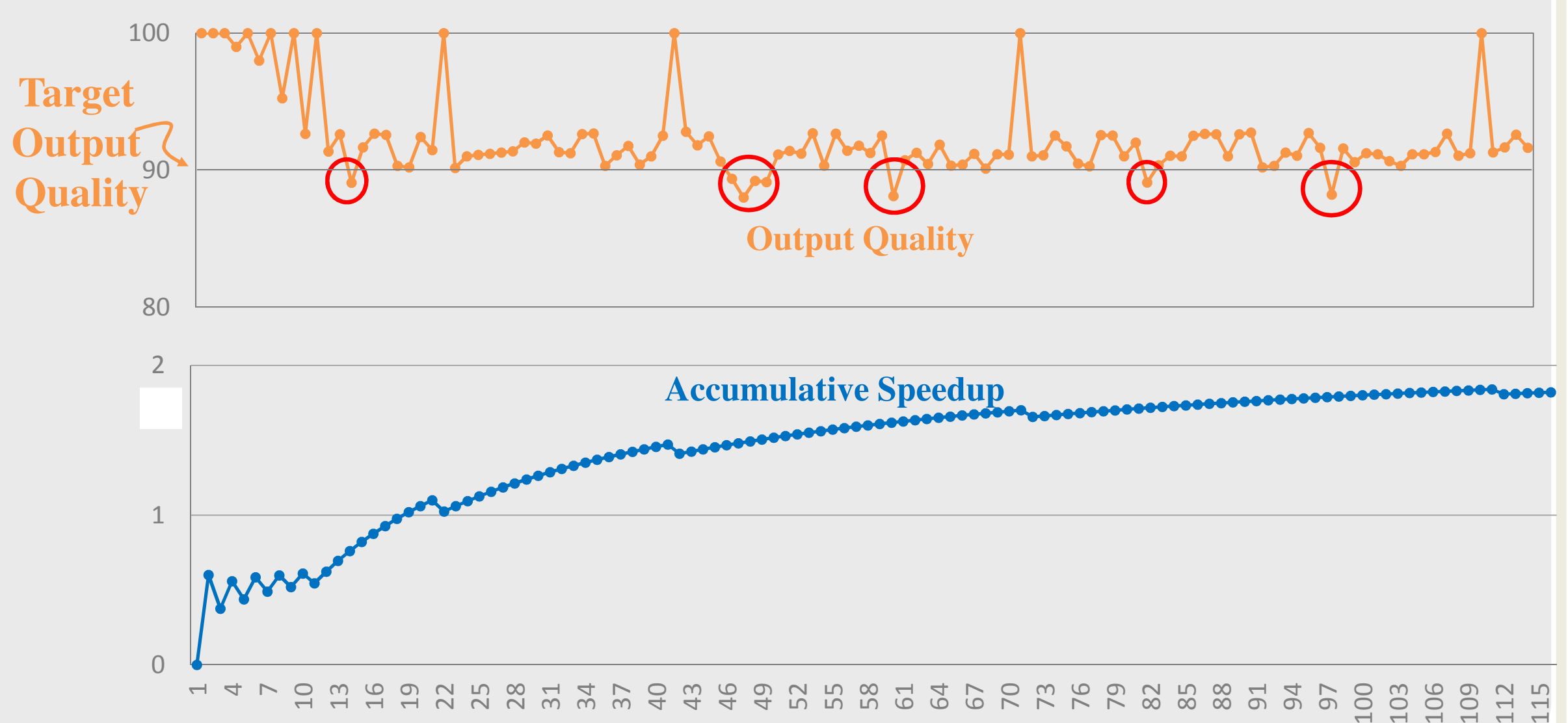


Tuning

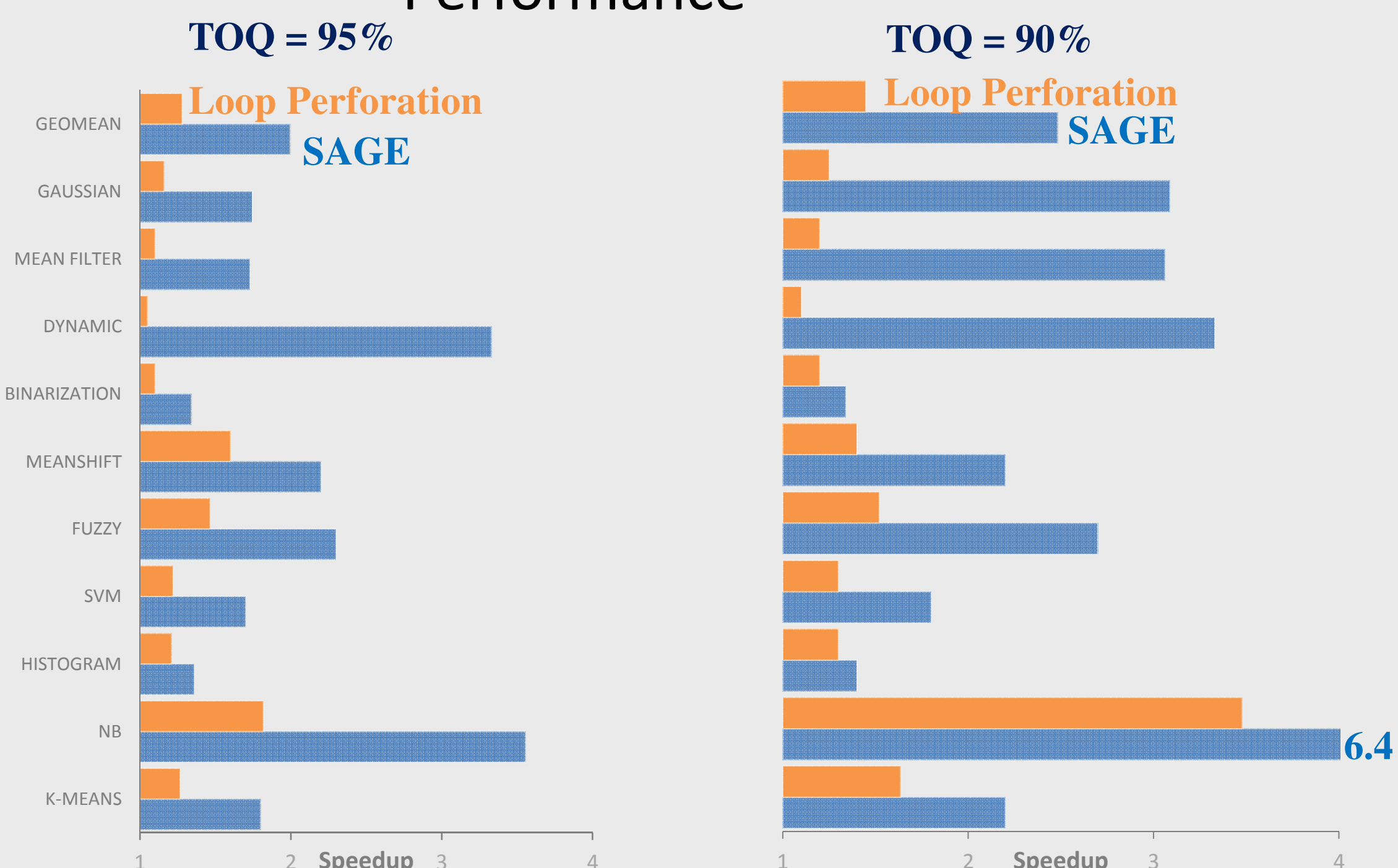
TOQ = 90%



K-Means Runtime



Performance



Conclusion

- SAGE enables the programmer to implement a program once
- It automatically generates approximate kernels with different parameters
- Runtime system uses tuning parameters to control the output quality during execution
- 2.5x speedup with less than 10% quality loss compared to the accurate execution