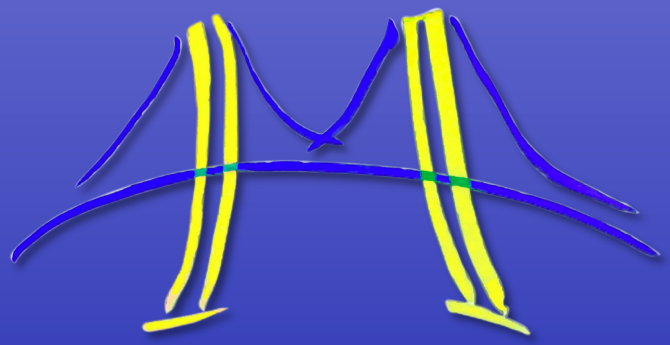
Communication-Avoiding QR for Computer Vision

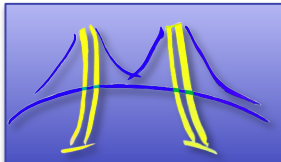
Michael Anderson

Grey Ballard

James Demmel

Kurt Keutzer





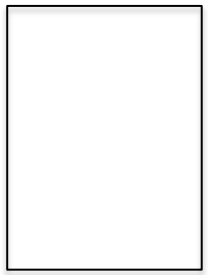
Introduction

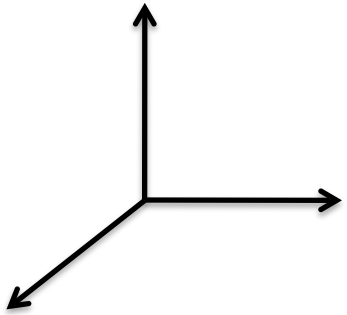
- The BEBOP group studies and designs cutting-edge parallel linear algebra algorithms
  - Communication-Avoiding QR
- The PALLAS group applies linear algebra to computer vision, to enable compelling applications
  - Robust video background subtraction
- Cutting-edge linear algebra enables cutting-edge computer vision!
  - (Cutting edge detection)

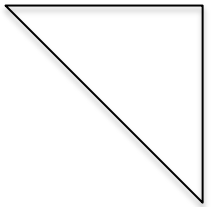
QR Decomposition

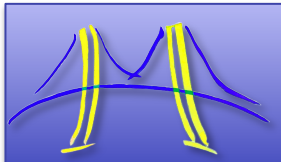
- Factor a matrix A into an orthogonal matrix Q, and an upper triangular matrix R

A = Q * R









Applications of QR

- Least squares → Linear Regression

```
>> [Q R] = qr(A, 0)
>> w = R\(Q'*b)
```

- Communication-avoiding Krylov methods → Sparse linear solvers
  - See Nick Knight's talk tomorrow
- Cleaning your room?

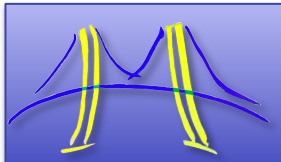
Cleaning Room with QR

After 30 QRs

After 70 QRs

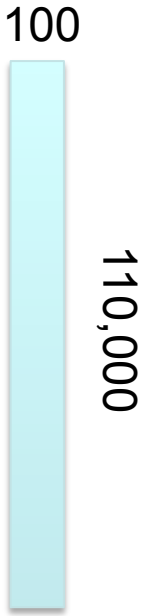
After 100 QRs

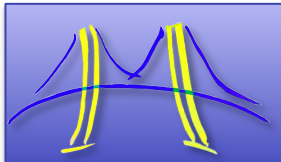
After 300 QRs



QR on Parallel Architectures

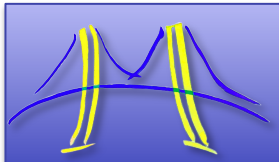
- "Isn't dense linear algebra on multicore/GPU a solved problem?"
  - NO!
- Consider QR decomposition of an entirely realistic tall-skinny matrix (e.g. 110000 x 100)
  - MAGMA (Fermi)
    - 12 Gflops. 0.9% of peak.
  - CULA (Fermi)
    - 5.1 Gflops. 0.4% of peak.
  - MKL (Dual Nehalem)
    - 15 Gflops. 10% of peak.





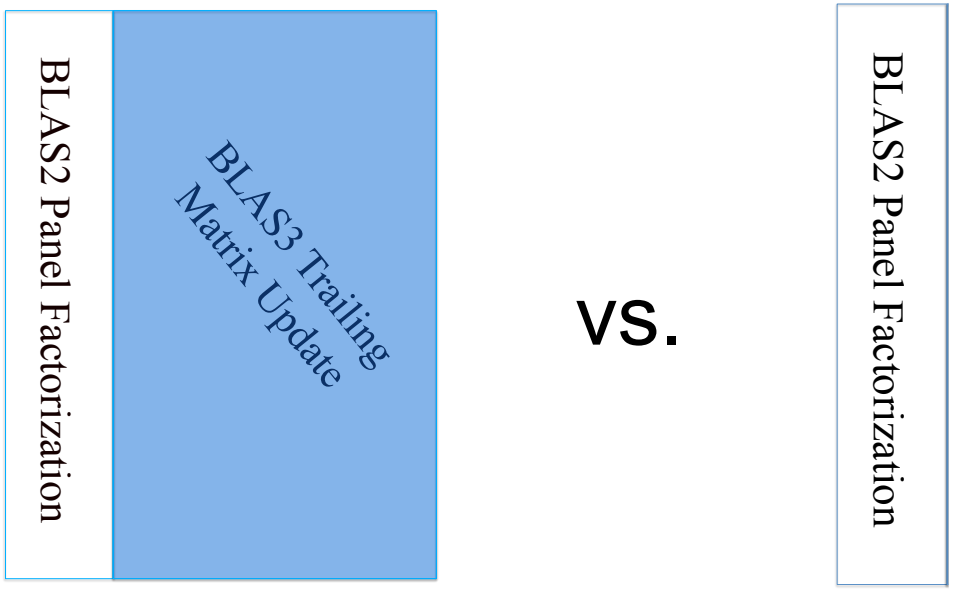
QR on Parallel Architectures

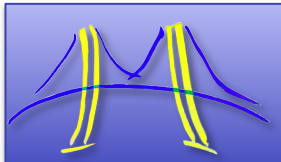
- Tall-skinny QR is actually common
  - Communication-avoiding Krylov methods
    - Dimension of sparse matrix (e.g. 1 million) x Number of steps (e.g. 10)
  - The “cleaning room” example (robust background subtraction)
    - Number of pixels (e.g. 110,000) x number of frames (e.g. 20)
  - Least squares
    - When the number of observations is much greater than the number of variables
    - e.g. Boston Housing Dataset is 506 houses x 14 variables



Householder Algorithm for QR

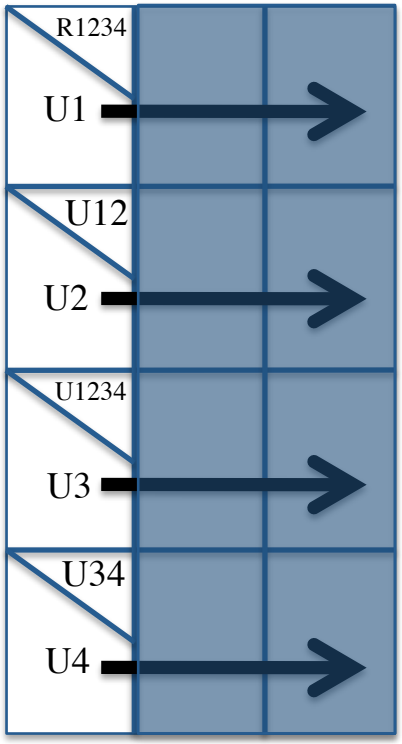
- Why is the tall-skinny case slow?
  - BLAS2 == Bandwidth Bound
  - "Communication" is heavy between DRAM and processor cores

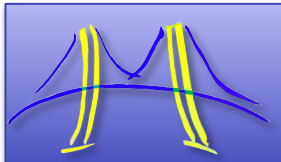




Communication-Avoiding QR

- Communication-Avoiding algorithm breaks the problem into small (cacheable) independent (parallel) problems
- Ideal for GPUs
- Communication-avoidance means fewer DRAM transfers
- Theoretically compute-bound



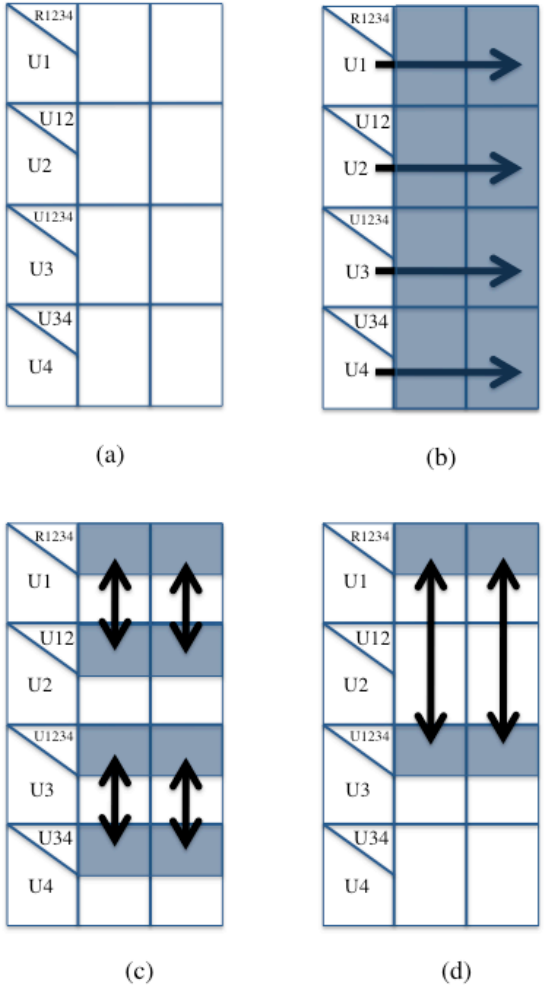


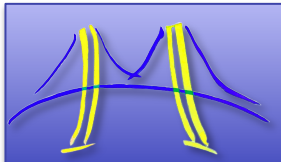
Why is this difficult on the GPU?

Must perform complete QRs within a thread block

Reduction tree must eliminate **irregular** triangles

Locations to be updated are **dispersed** throughout the matrix

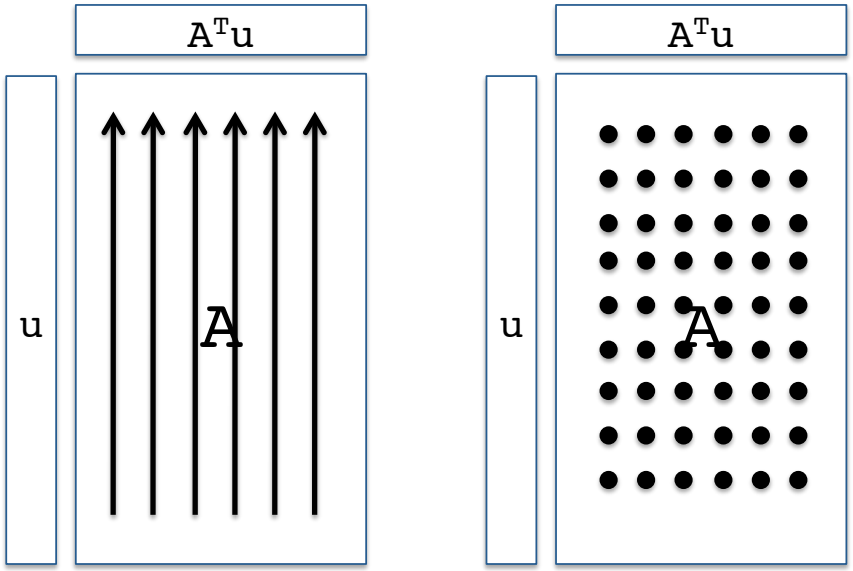




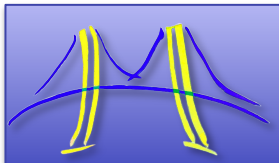
CAQR Autotuning

- Segmented reduction (matrix-vector multiply) and rank-1 update are the core computations
  - Referenced from all kernels
- Use this snippet to determine optimal block size

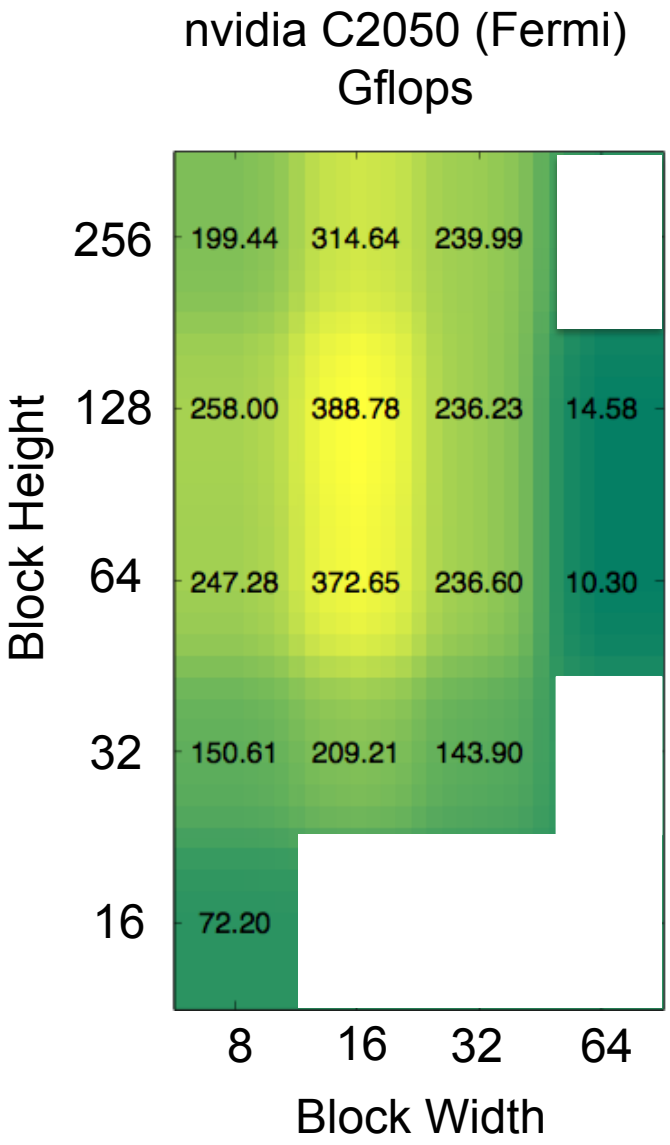
```
// For each Householder vector
 for(int j = 0 ; j < BLK_WIDTH; j++) {

   // Matrix-vector multiply
   real_t res = reduce(u_sh, col, av, tid_u, tid_l, tid);

   // Rank-1 update
   update(u_sh, col, res, tid_u);

   // Go to the next Householder vector
   u_sh += BLK_HEIGHT;
 }
```

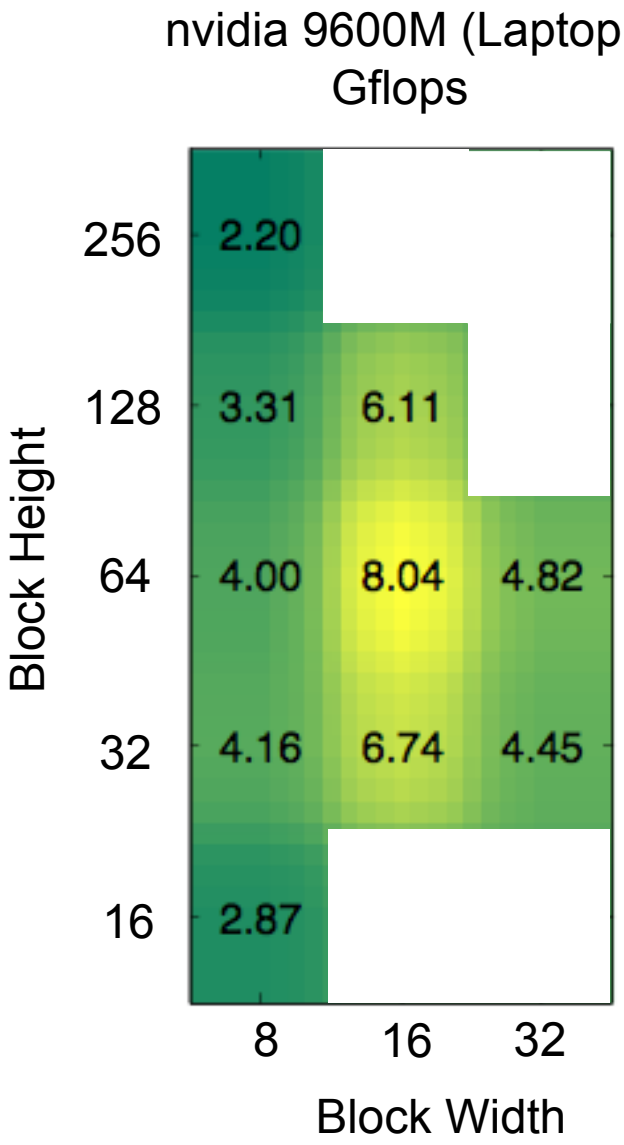


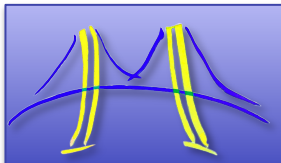
(a) (b)



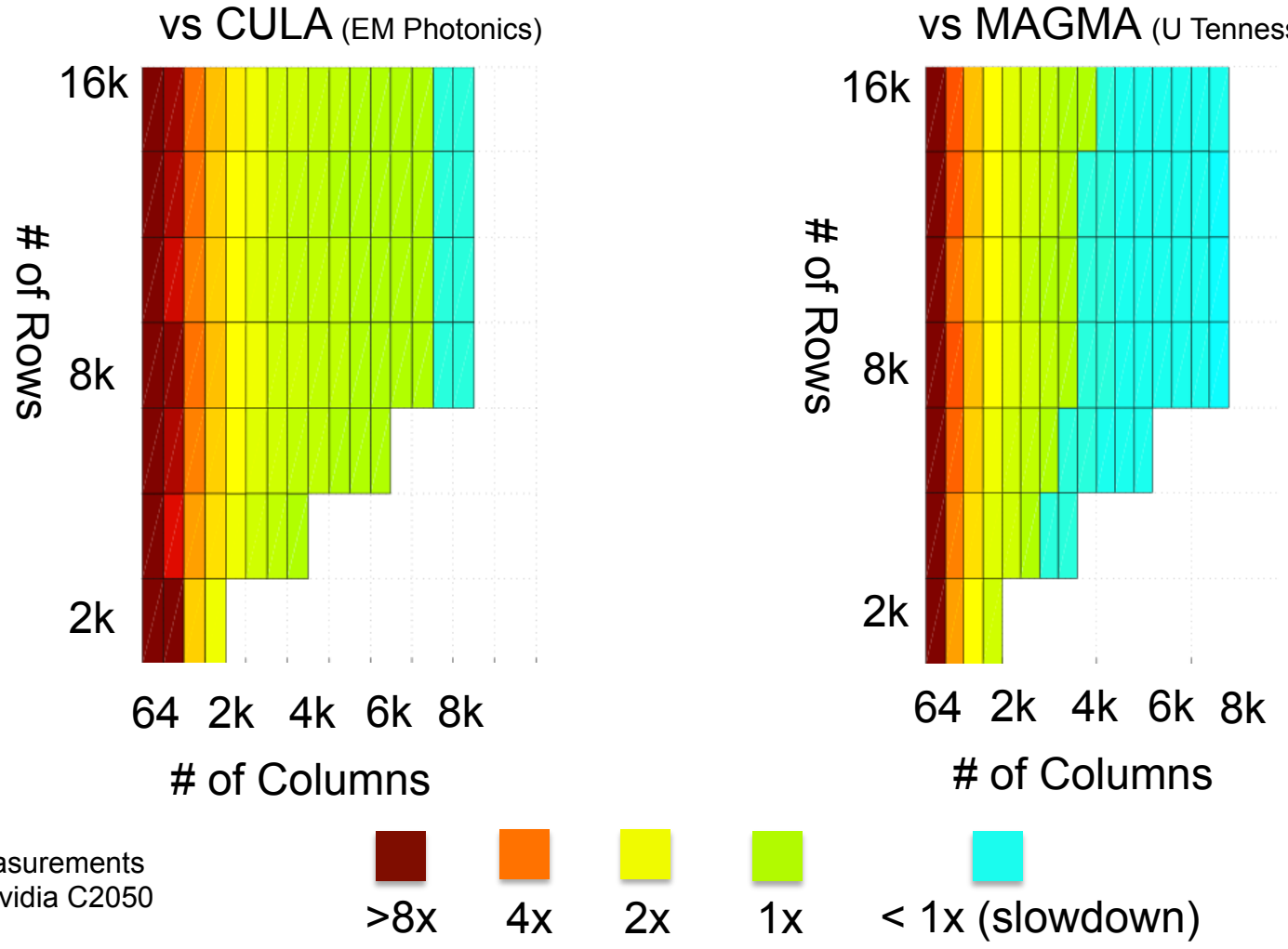
CAQR Autotuning



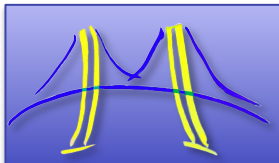




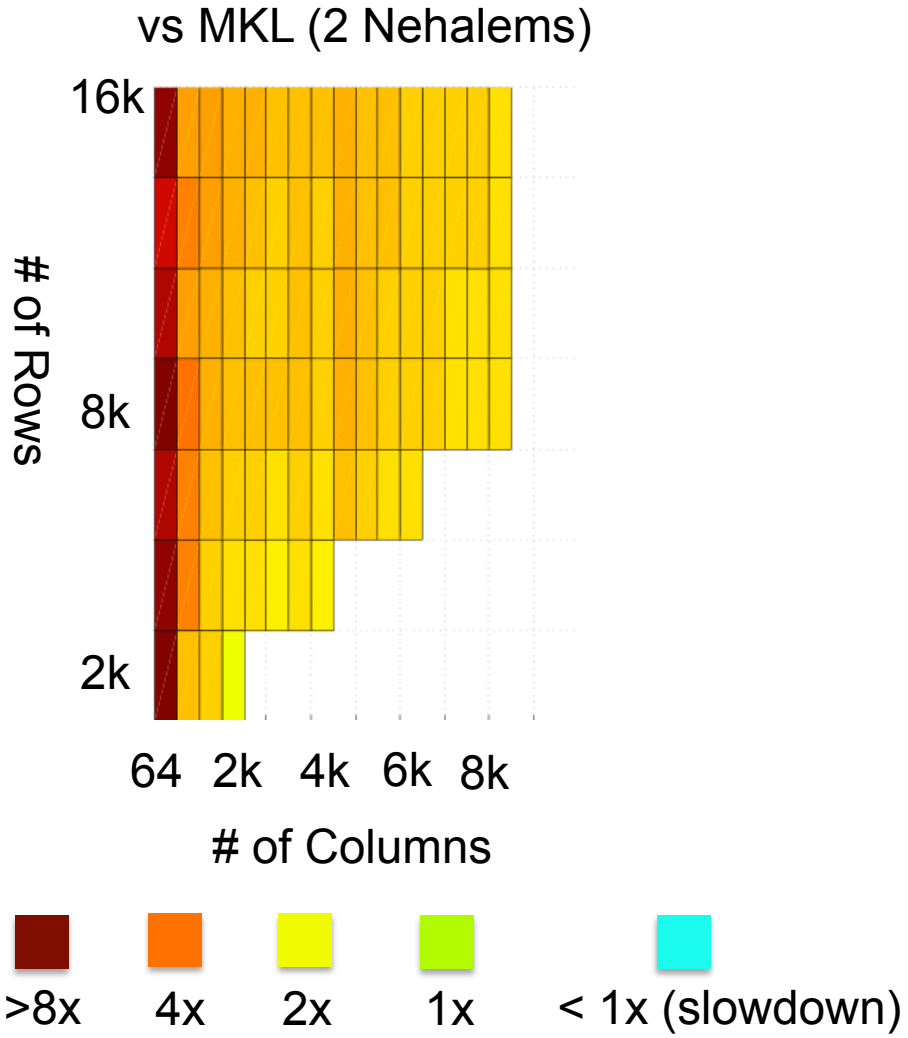
CAQR Speedup vs GPU Libraries



”Communication-Avoiding QR for GPUs" Anderson, Ballard, Demmel, Keutzer IEEE International Parallel & Distributed Processing Symposium, 2011



CAQR Speedup vs MKL



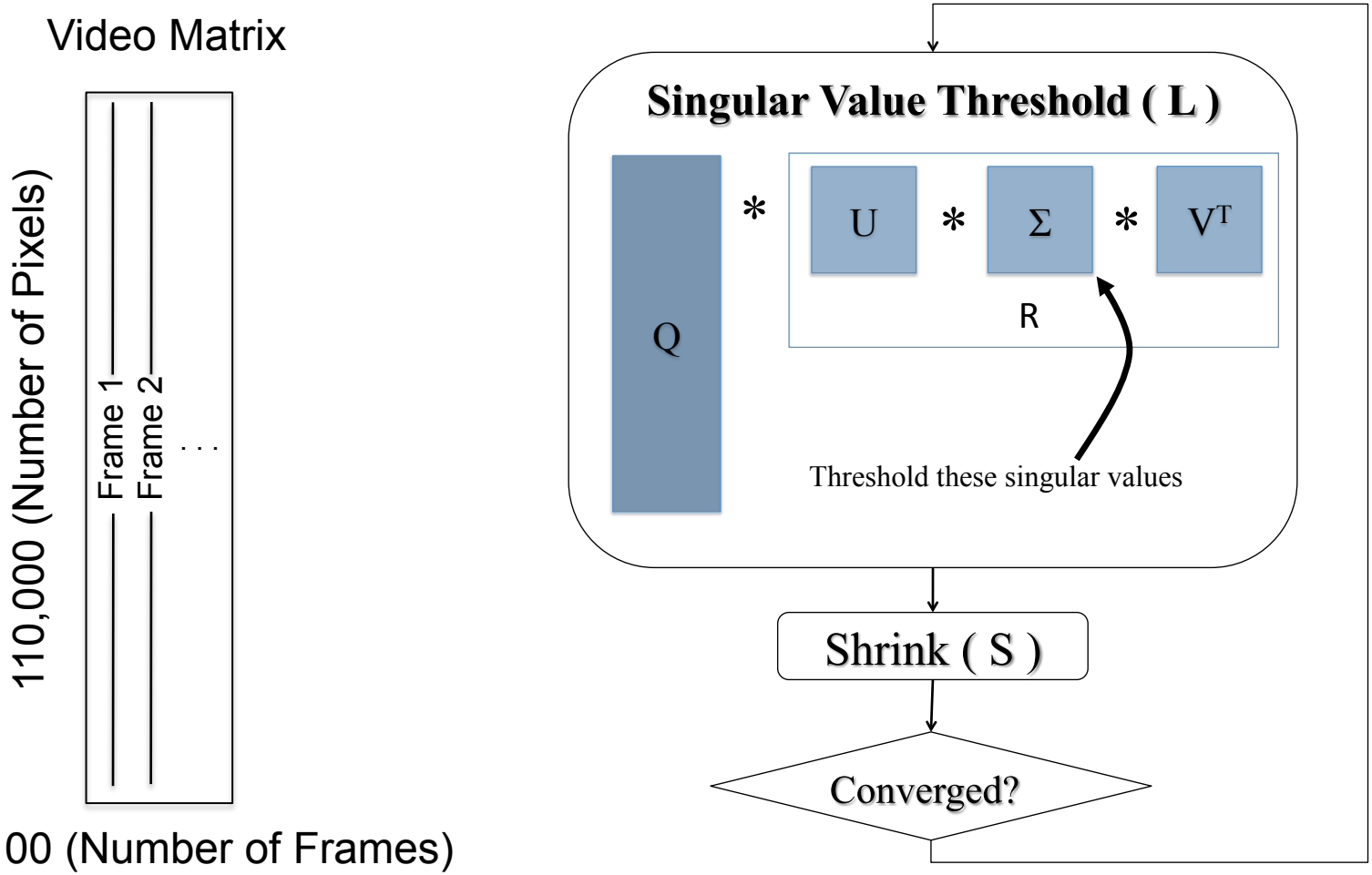
”Communication-Avoiding QR for GPUs" Anderson, Ballard, Demmel, Keutzer IEEE International Parallel & Distributed Processing Symposium, 2011

Robust Background Subtraction

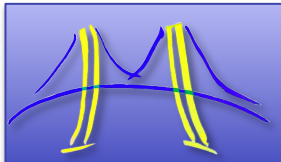
- Decompose a video into a low-rank component (background) and a sparse component (foreground)
- Background is the *principal component*

From “Robust Principal Component Analysis?” by Emmanuel Candes et al

Robust Background Subtraction



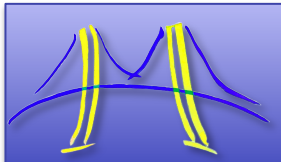
From “Robust Principal Component Analysis?” by Emmanuel Candes et al



Robust Background Subtraction

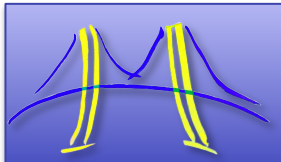
SVD type	Number of Iterations/Sec.
MKL SVD (4 cores)	0.9
BLAS2 QR (GTX480)	8.7
CAQR (GTX480)	27.0

- Tech Transfer: Others are using our QR software
  - *Y. Dong, G. N. DeSouza. "Adaptive learning of multi-subspace for foreground detection under illumination changes". Journal of Computer Vision and Image Understanding, 2010.* (Not in paper, but in future public releases)
  - Planning to integrate into Trilinos large-scale linear algebra library (Sandia)
  - Planning to integrate into MAGMA (U Tennessee)



Summary

- QR Decomposition is very useful
  - Least Squares, Communication-avoiding Krylov methods, surveillance video
- Communication-Avoiding QR is a great fit for GPU/ Multicore
  - Especially for tall-skinny matrices
  - Up to 13x faster than other **parallel** libraries
- Tall-skinny QR is at the core of robust video background subtraction
  - 30x speedup (parallel vs. parallel) for computer vision
- See our upcoming paper "Communication-Avoiding QR for GPUs" in IPDPS 2011



Thank You

- Michael Anderson
  - mjanders@eecs.berkeley.edu
- Grey Ballard
  - ballard@cs.berkeley.edu
- Jim Demmel
- Kurt Keutzer
- This is just one small piece. See Nick Knight's talk tomorrow for more on Communication-Avoiding algorithms