QR with column pivoting in PLASMA

ICL Lunch Talk, Feb 15, 2013

Mark Gates

Outline

- Motivation
- QP3 algorithm
- Challenges in PLASMA Implementation
- Current Results
- Future Directions

Motivation

- QR is backward stable, without pivoting
  - Meaning backward error is small,

$$\frac{\|\boldsymbol{A} - \boldsymbol{Q}\boldsymbol{R}\|}{\|\boldsymbol{A}\|} = O(\varepsilon_{\text{machine}})$$

- But if $\boldsymbol{A}$ is rank-deficient, then $\boldsymbol{R}$ is singular

Example 1

```
>> A = [ v1, v2, (v1+v2), v3 ]
A =      1     5      6      4
         2     4      6      2
         3     3      6      8
         4     2      6      6
         5     1      6     10
```

linearly dependent column

zero on diagonal

```
>> [Q,R] = qr( A );
>> R
R =   -7.4162   -4.7194  -12.1356  -14.2930
           0   -5.7208   -5.7208   -1.1442
           0         0   -0.0000    1.9388
           0         0         0    3.2621
           0         0         0         0
```

Example 2

bad scaling

linearly dependent column

```
>> A = [ v1, eps*v2, (v1+v2), v3 ]
A =       1        5        6        4
          2        4        6        2
          3        3        6        8
          4        2        6        6
          5        1        6       10
```

zeros on diagonal

```
>> [Q,R] = qr( A );
>> R
R =   -7.4162   -0.0000  -12.1356  -14.2930
           0   -0.0000   -5.7208   -1.1442
           0         0   -0.0000    1.0231
           0         0         0    3.6542
           0         0         0         0
```

Definition

- QR with column pivoting

$$\boldsymbol{AP} = \begin{bmatrix} \boldsymbol{Q}_1 & \boldsymbol{Q}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \boldsymbol{0} & \boldsymbol{R}_{22} \end{bmatrix}, \quad \boldsymbol{R}_{22} \approx \boldsymbol{0}$$

- Where:

  $\boldsymbol{A}$ has rank k

  $\boldsymbol{P}$ is permutation matrix

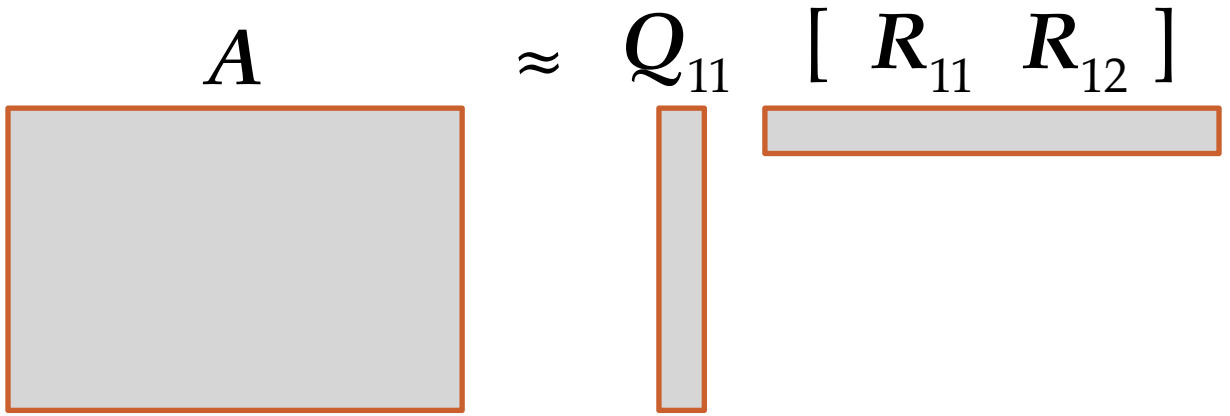
  $\boldsymbol{R}_{11}$ is k x k and well-conditioned

Uses

- Rank-deficient least squares, solve $\boldsymbol{Ax} = \boldsymbol{b}$

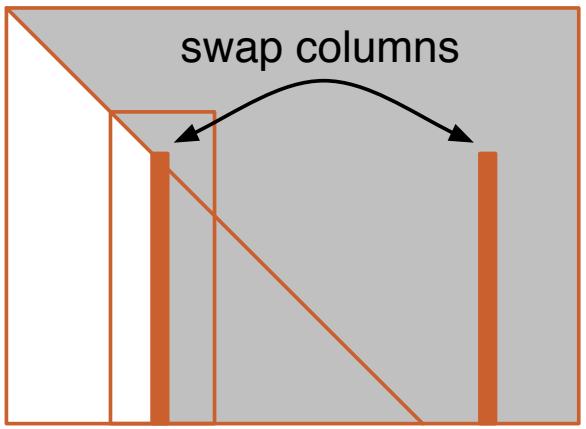
$$\boldsymbol{z} = \boldsymbol{R}_{11}^{-1}\boldsymbol{Q}_1^T\boldsymbol{b}, \qquad \boldsymbol{x} = \boldsymbol{P}\begin{bmatrix}\boldsymbol{z}\\ \boldsymbol{0}\end{bmatrix}$$

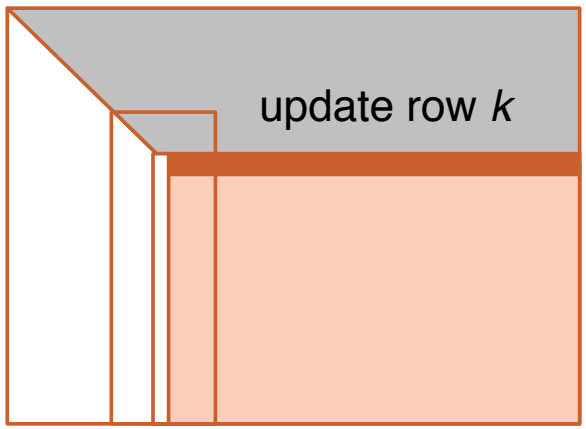
- Low-rank approximation

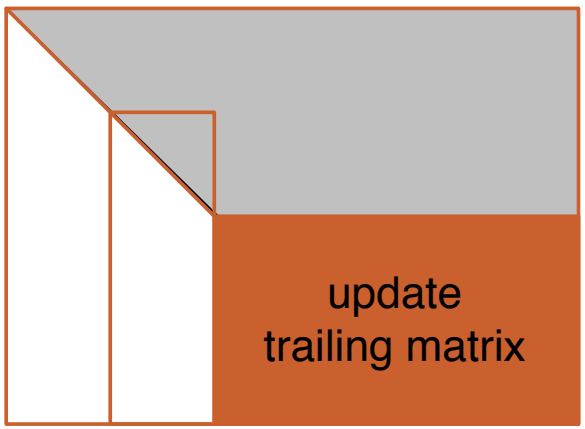


QP3 Algorithm

compute column norms
for each panel
    for each column k in panel
        pivot with column of largest norm
        factor column (Householder)
        update row k of trailing matrix — requires GEMV
        update column norms, subtracting row k (*)
    update trailing matrix — GEMM with trailing matrix

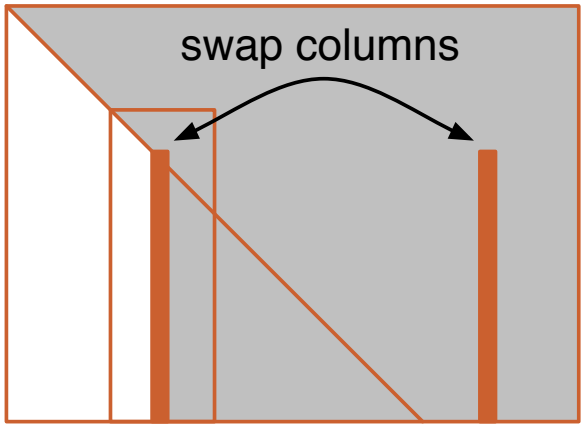


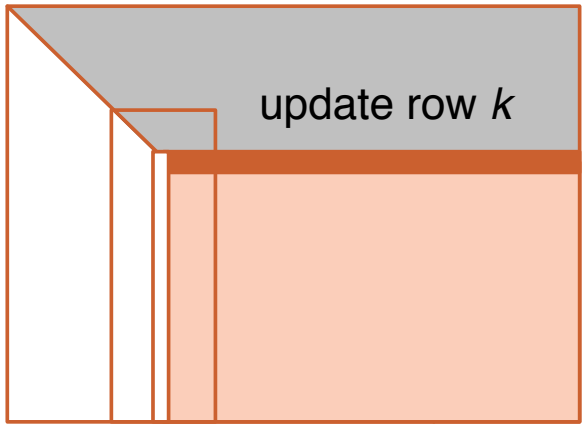


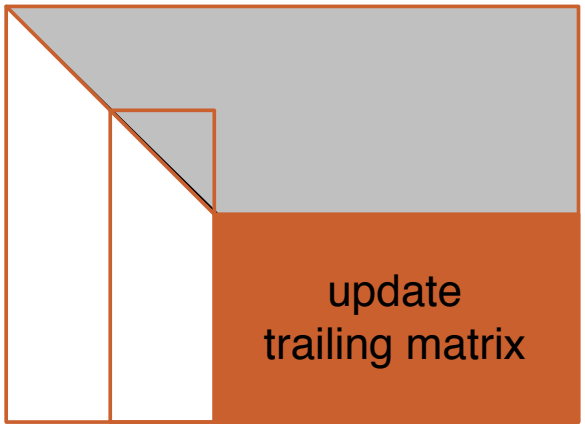


Challenges in PLASMA

- Entire column is factored, instead of tiles
- Pivoting forces synchronization
- GEMV starts mid-tile — dependency issues
- GEMV has run-time value tau — pass as pointer







Challenges

- Updating column norms can incur cancellation
- Example:

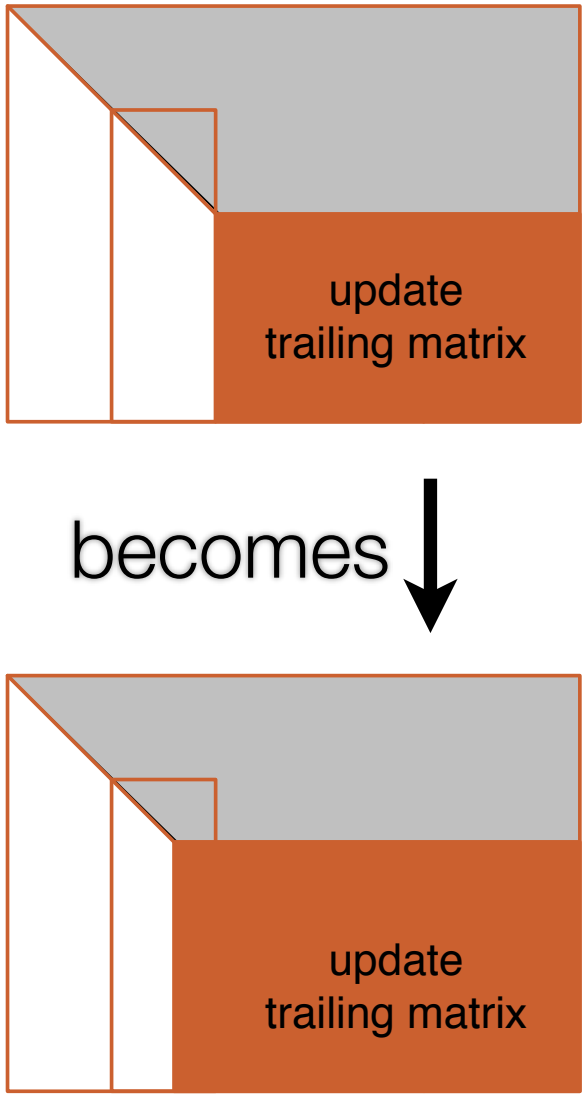
```
v = [  1
       1.23e-8
       1.23e-8 ]
x = norm(v)

x = sqrt( x^2 - v(1)^2 )
   = 2.1073e-08

norm( v(2:3) )
   = 1.7395e-08
```

Challenges

- Solution:
  - Stop panel
  - Update trailing matrix
  - Restart panel from column k
- Requires synchronized task execution
- Runtime decision:
  DAG cannot be known ahead

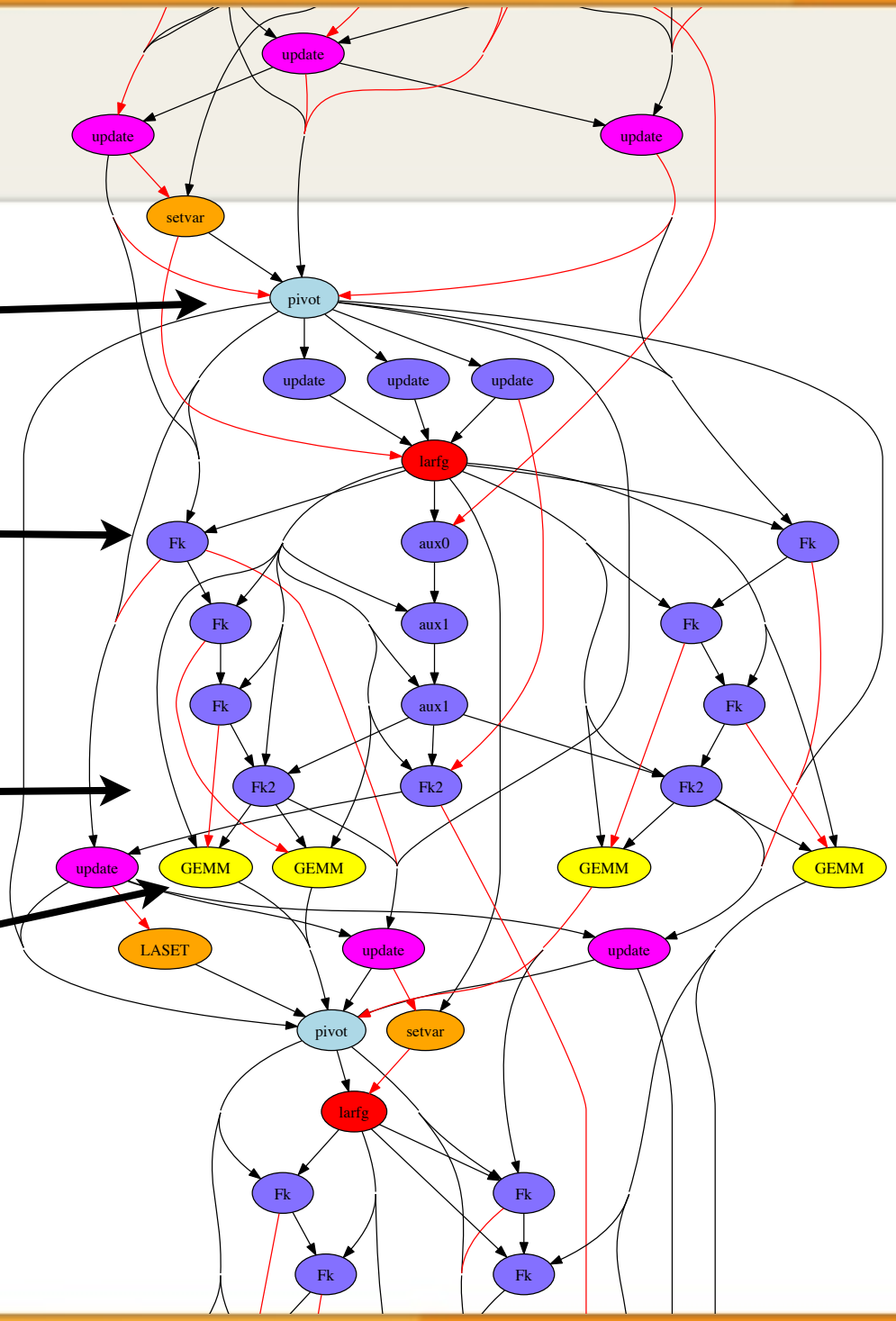


DAG

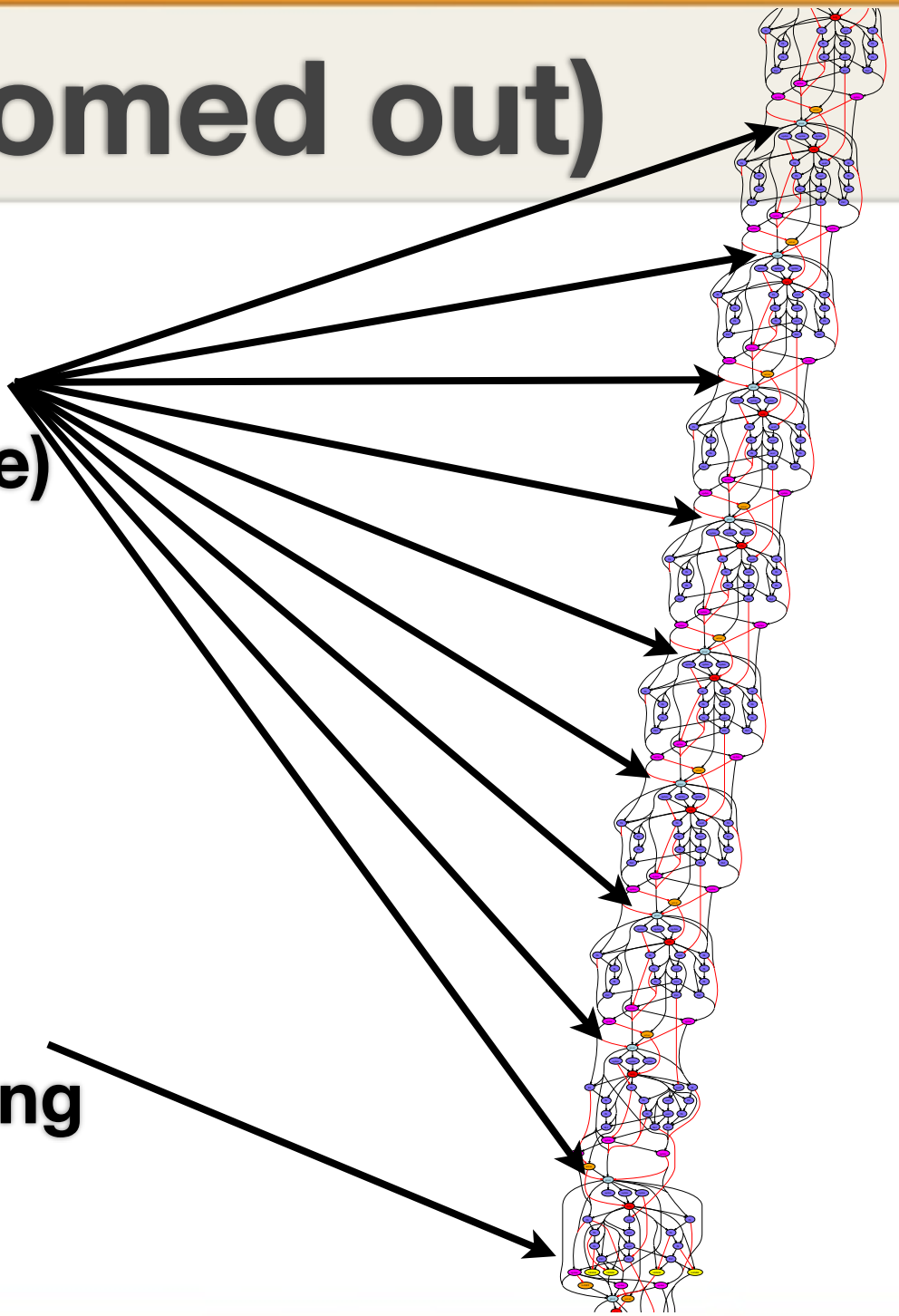
pivot (synchronize)

GEMVs (update row k)

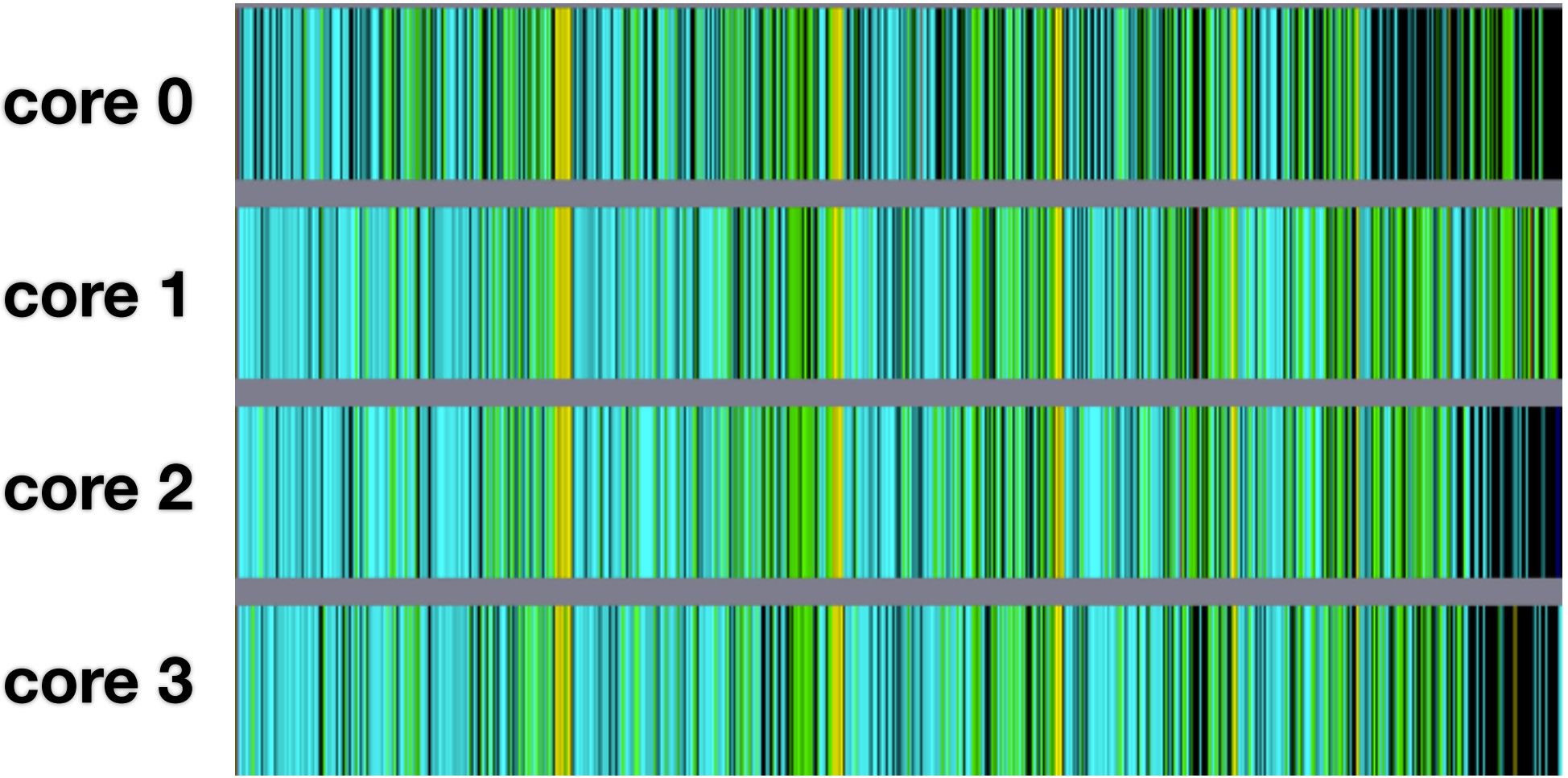
GEMM (update trailing matrix)



DAG (zoomed out)



Execution trace

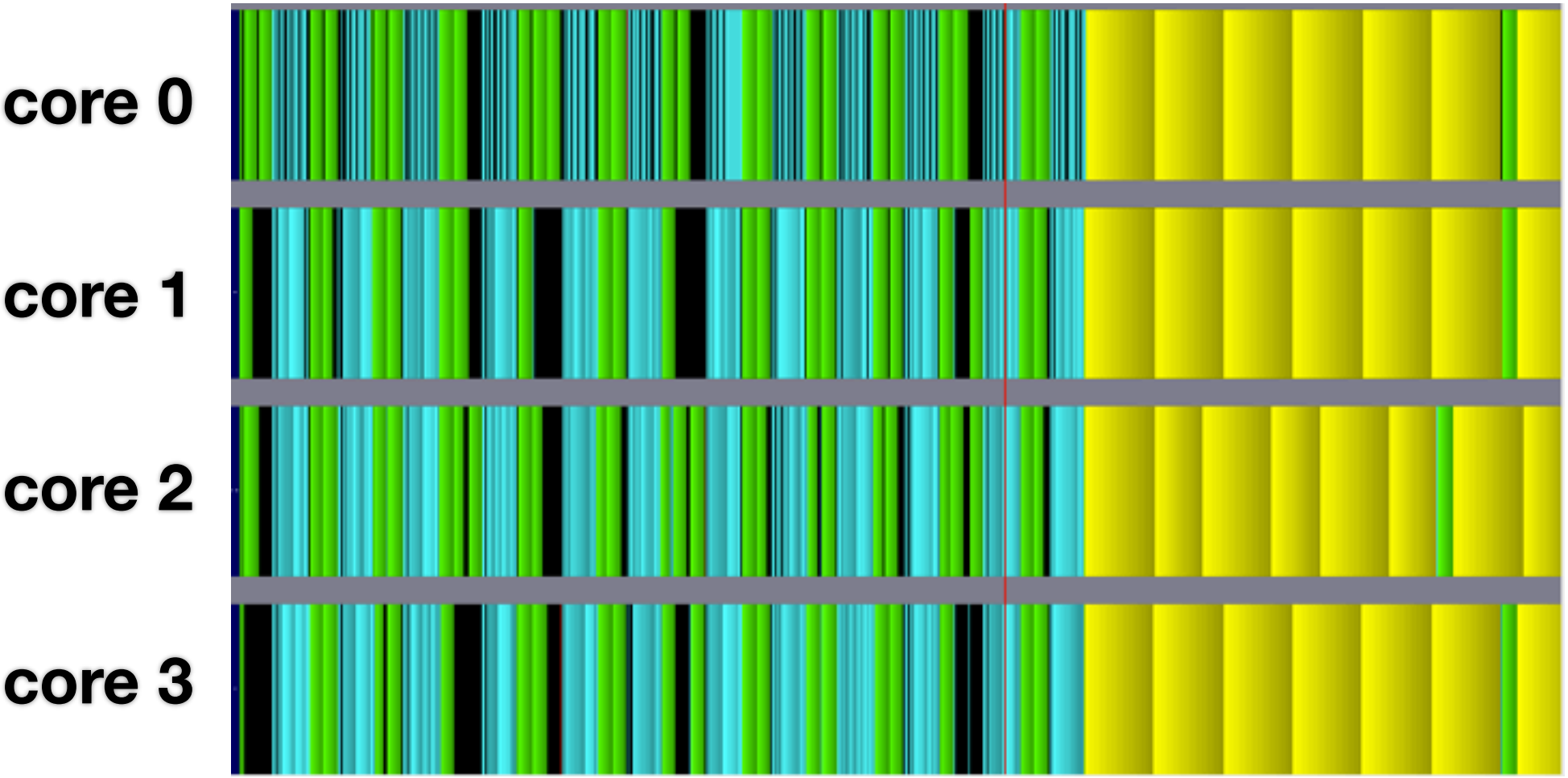


GEMV

update norms

GEMM

Execution trace



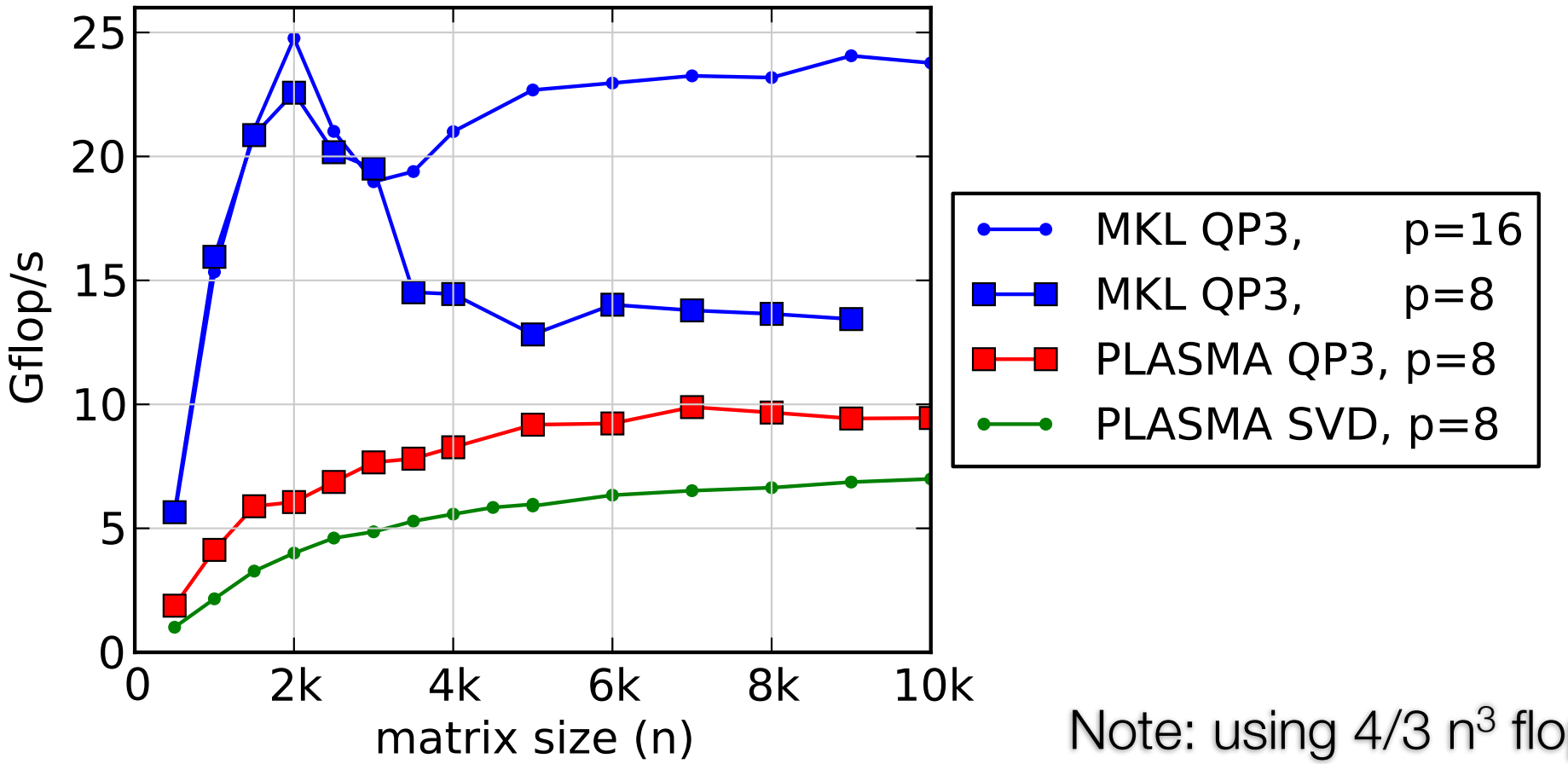
GEMV

update norms

GEMM

Current Results

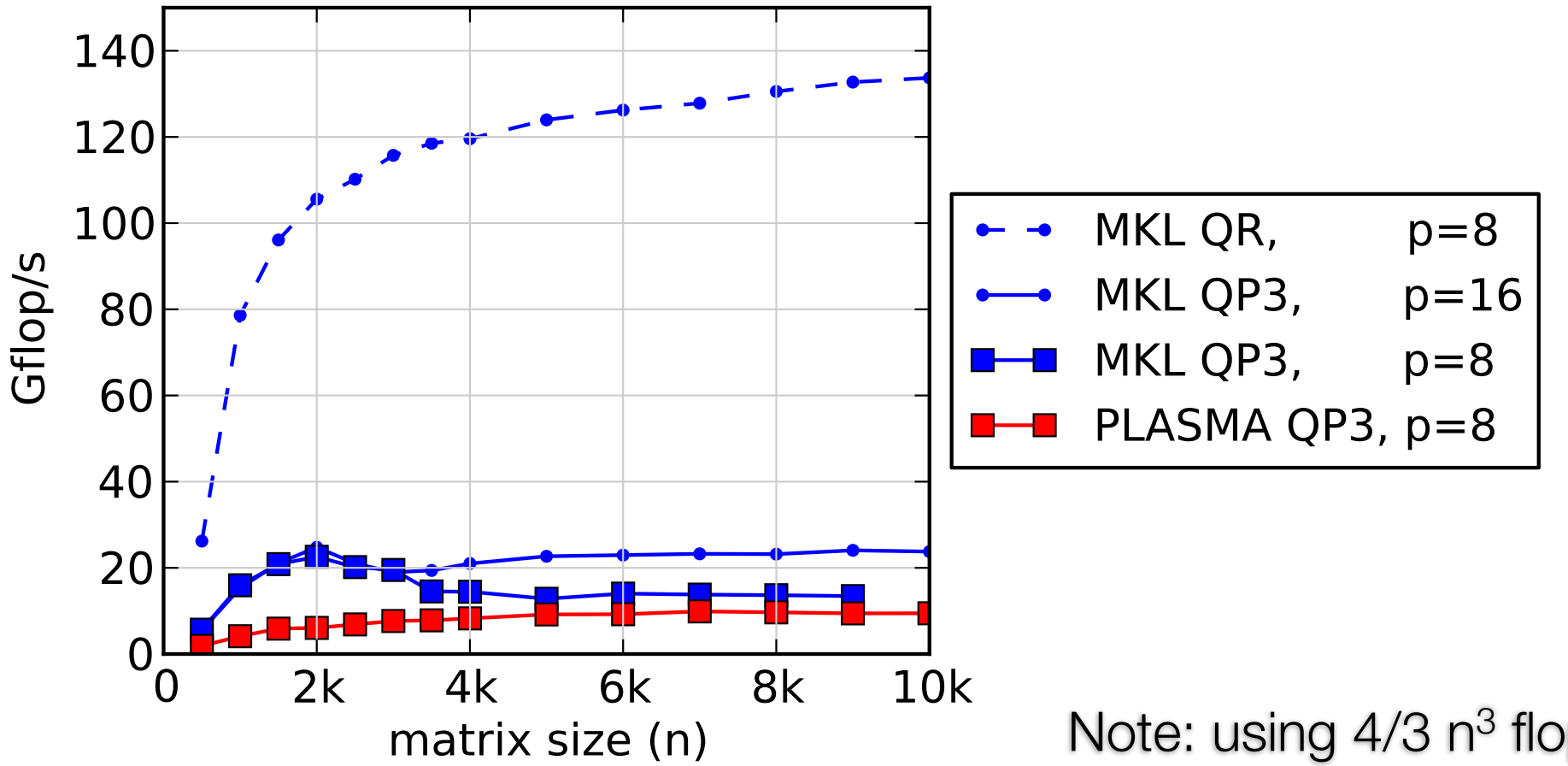
- On bunsen (2x8 core Sandy Bridge, 2.6 GHz)



Note: using 4/3 n^3 flops for QP3 and SVD

Current Results

- On bunsen (2x8 core Sandy Bridge, 2.6 GHz)



Note: using 4/3 n^3 flops for QP3 and SVD

Future Work

- Optimize PLASMA implementation
  - Eliminate redundant dependencies
  - Merge small tasks together
  - Static scheduling
- Examine RRQR (Rank Revealing QR) algorithm
  - Windowed pivoting, instead of whole trailing matrix
  - Post-process $\boldsymbol{R}$ using condition estimates