ABFT in Dense Linear Algebra

Yulu Jia, George Bosilca, Piotr Luszczek, Jack Dongarra

SIAM PP
Portland, OR
Feb 19, 2014

Motivation

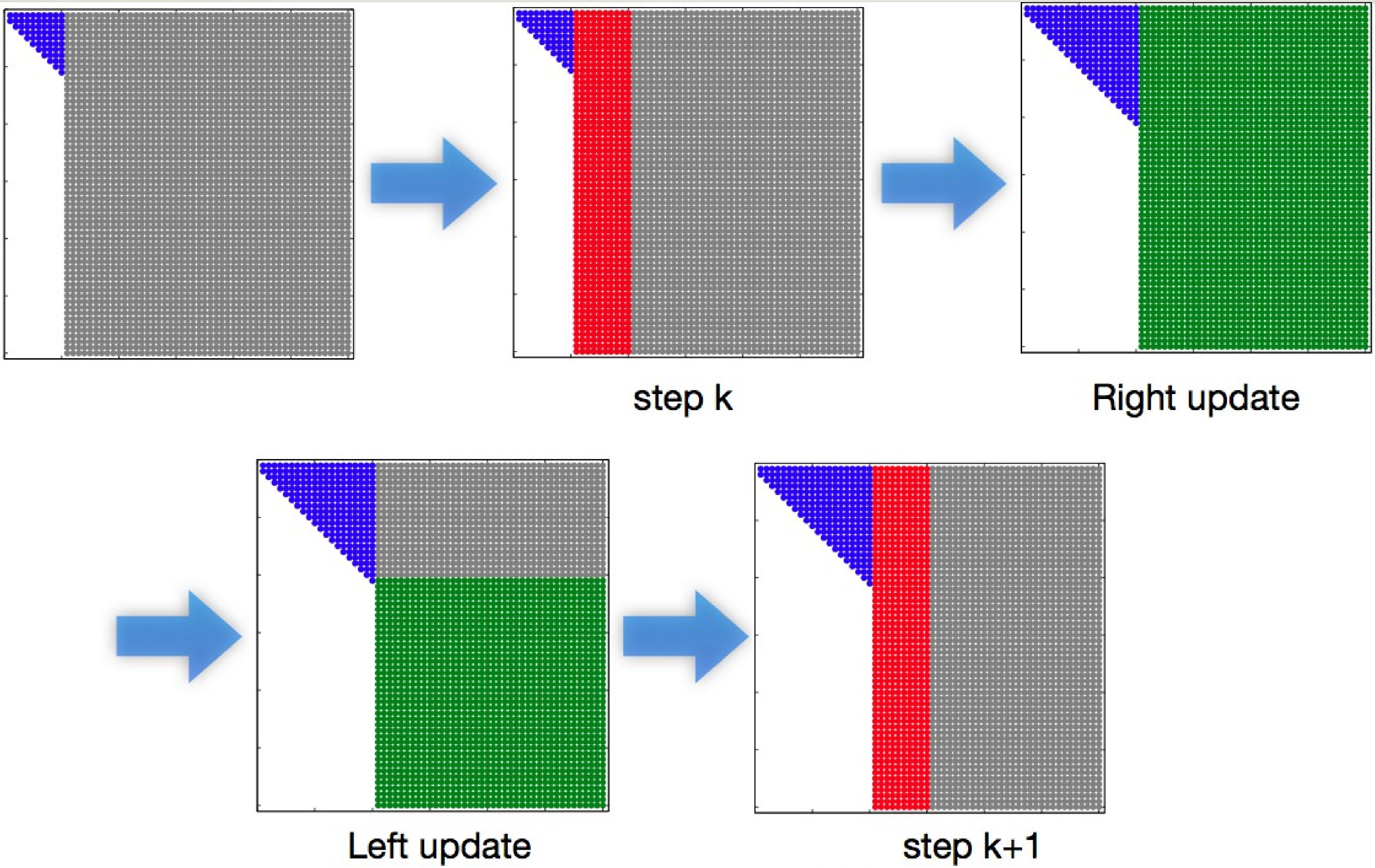
- Soft errors…
  - caused by: cosmic rays (alpha particle, high energy and/or thermal neutrons)
  - occur in practice
    - Commercial study in 2000 by Sun Microsystems
    - ASC Q supercomputer at Los Alamos in 2003
    - Jaguar (Cray XT5) at ORNL
      - Nearly 225k cores
      - 1253 separate node crashes during 537 days (Aug 2008-Feb 2010)
      - Or 2.33 failures per day Or less 10 hours of failure-free operation
    - … and any non-ECC machine
- Accelerators are common
  - In many shared-memory systems
  - Supercomputers
    - Tianhe-1A, Titan (Cray XK7, 560k cores), Tianhe-2 (3M+ cores)
- And at Exascale $\sim$1 billion threads and MTTF $<$ 1 day!

Importance of a Single Node

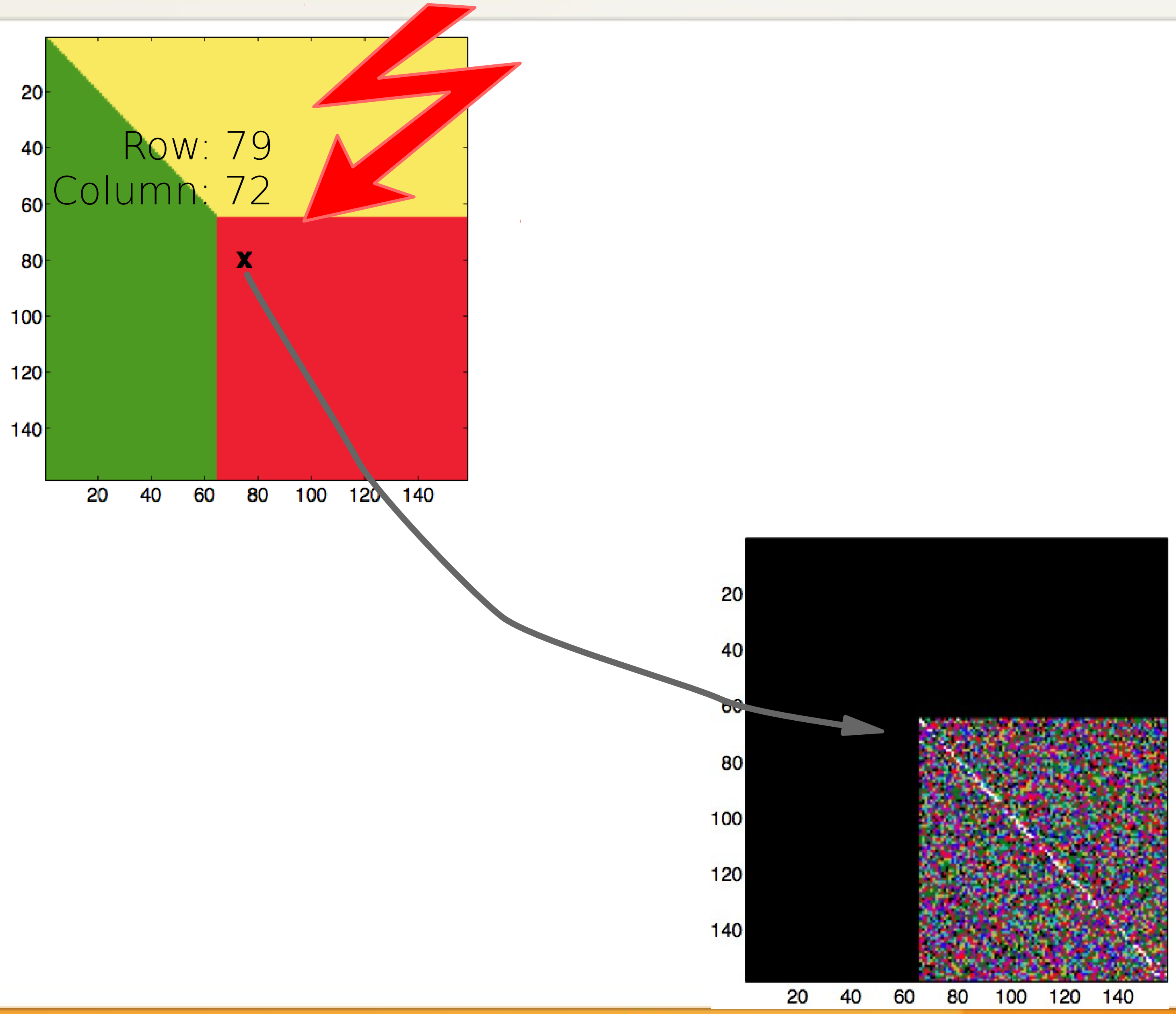
- MTTF of the entire machine depends on reliability of each node
- The MTTF of the entire machine can be statistically computed based on single-node reliability for a number of distributions
  - Exp(1/100)
  - Weibull(0.7, 1/100)
  - Weibull(0.5, 1/100)
- See Yves Robert's work for detailed analysis

One node $\sim 10^3$ cores	MTTF = 1 year	MTTF = 10 years	MTTF = 120 years
↓	↓	↓	↓
Exascale machine $\sim 10^6$ nodes	30 seconds	5 minutes	1 hour

Two-Sided Matrix Factorization



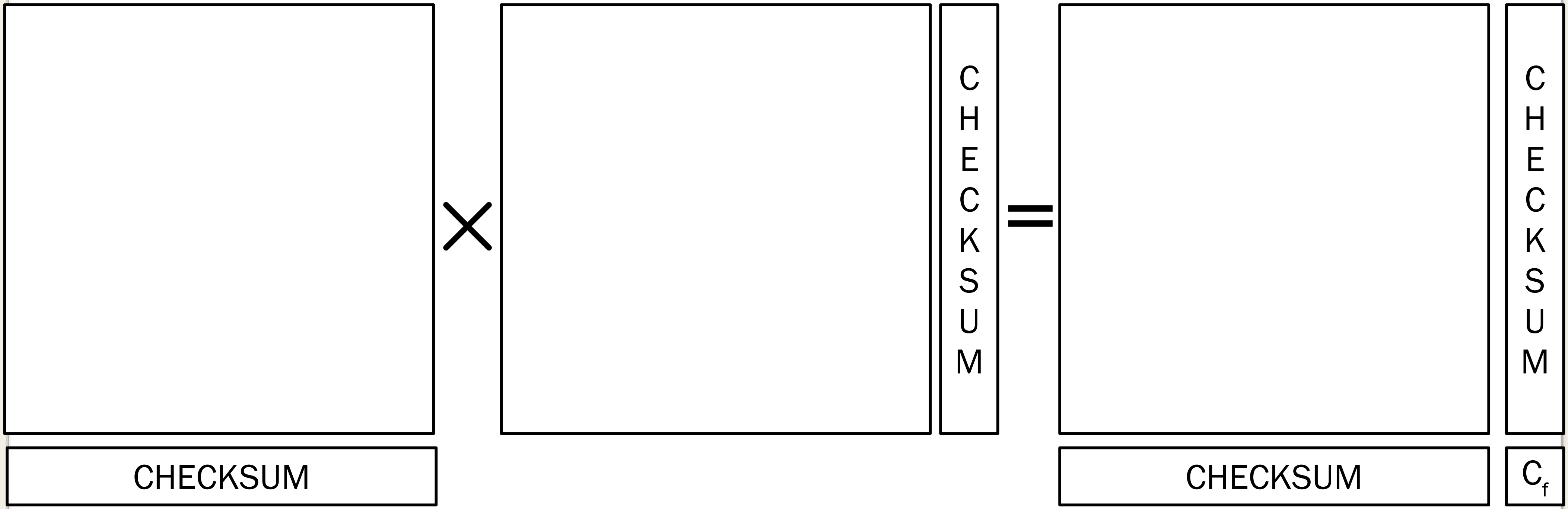
Error Propagation



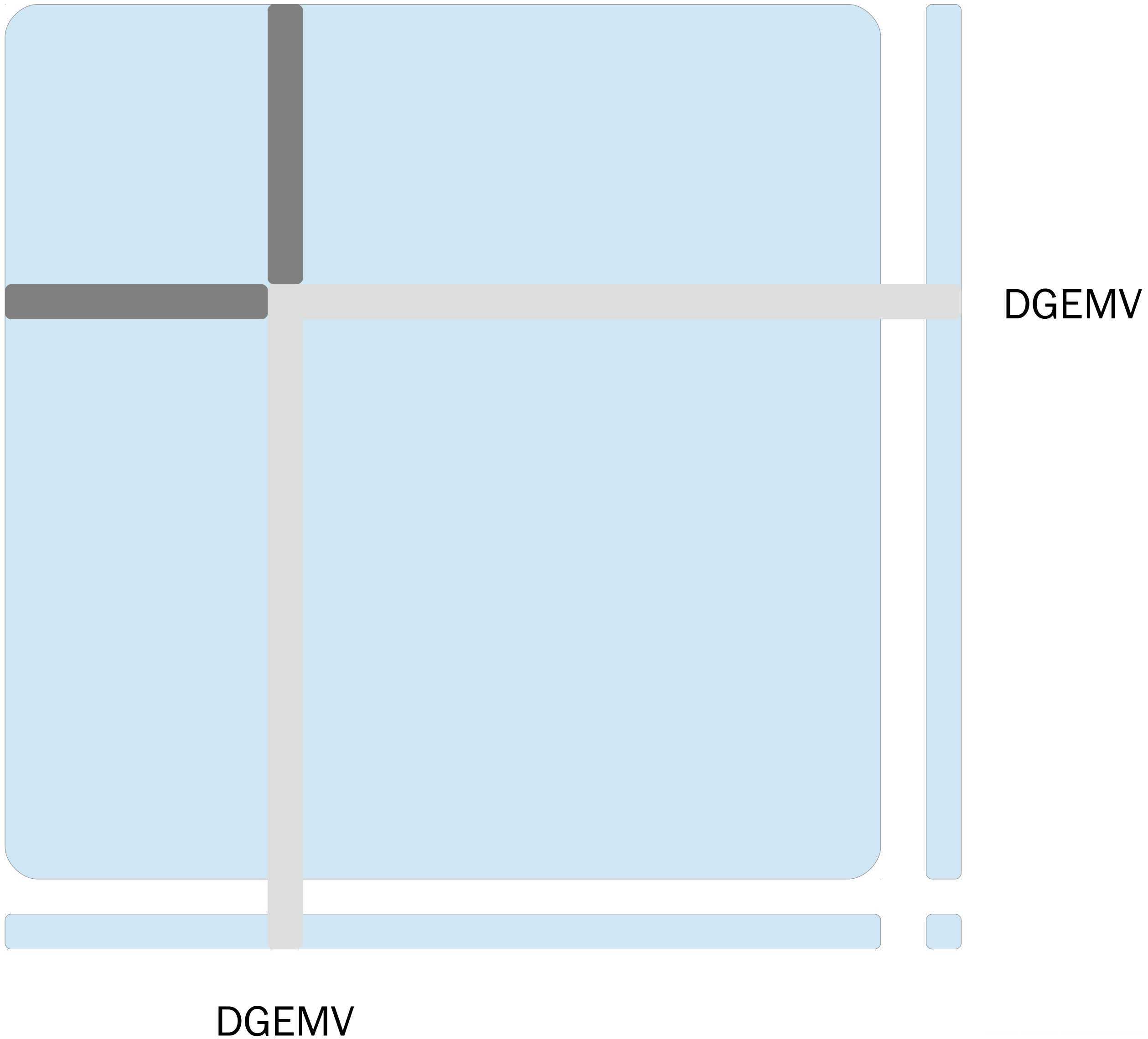
Techniques for Error Protection and Failure Recovery

- Algorithm-Based Fault Tolerance
  - Kuang-Hua Hua, Jacob Abraham, ABFT for Matrix Operations
    - Implementation on systolic arrays
  - Takes advantage of additional mathematical relationship(s)
    - Already present in algorithm
    - Introduced (cheaply, if possible) by ABFT – usually weighted sums
- Diskless checkpointing
  - Additional (small) data is kept in live processes
  - No need for full I/O checkpointing

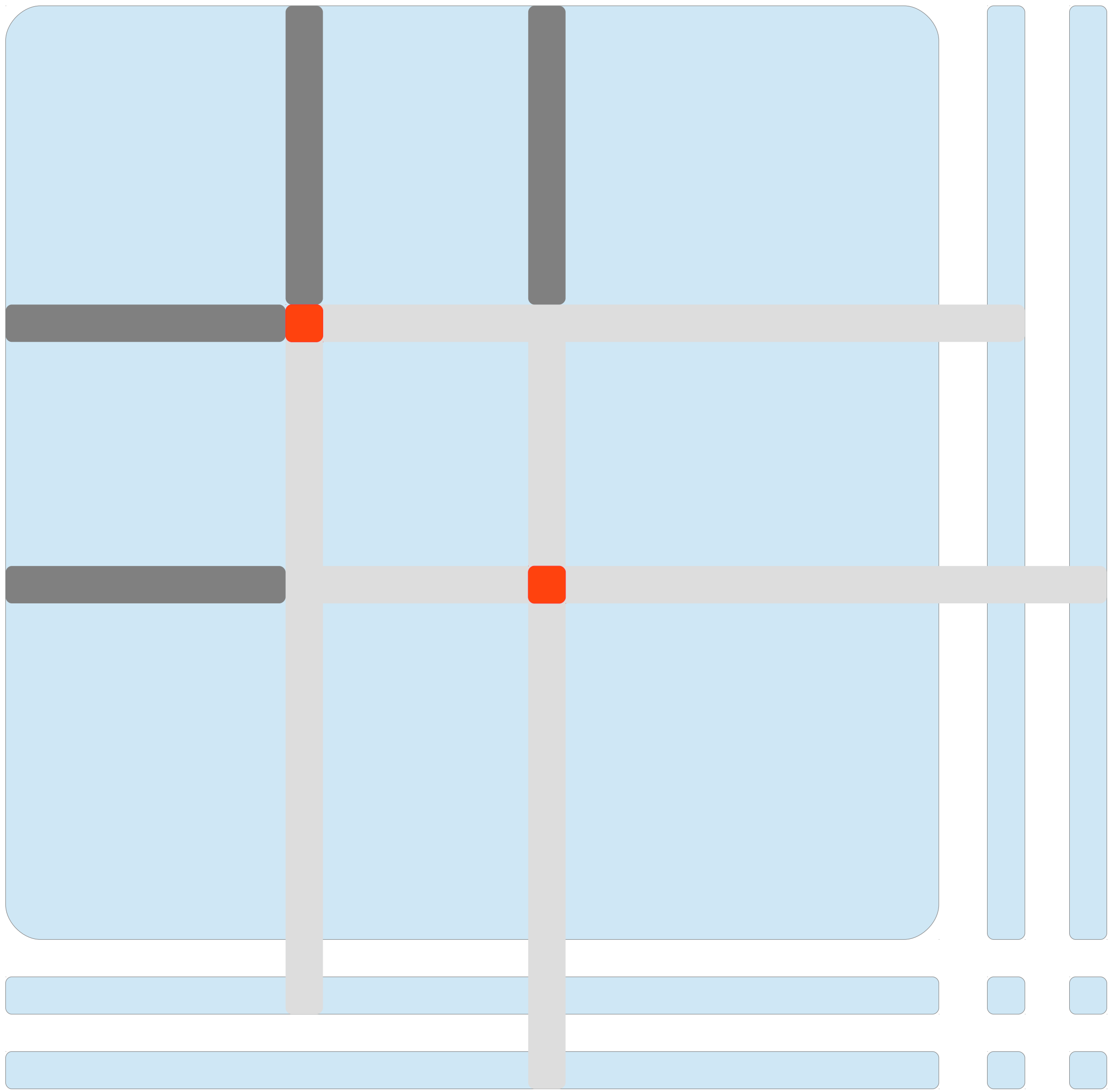
Huang&Abraham: Checksum M-M-mul



Extra Computation for Error Protection

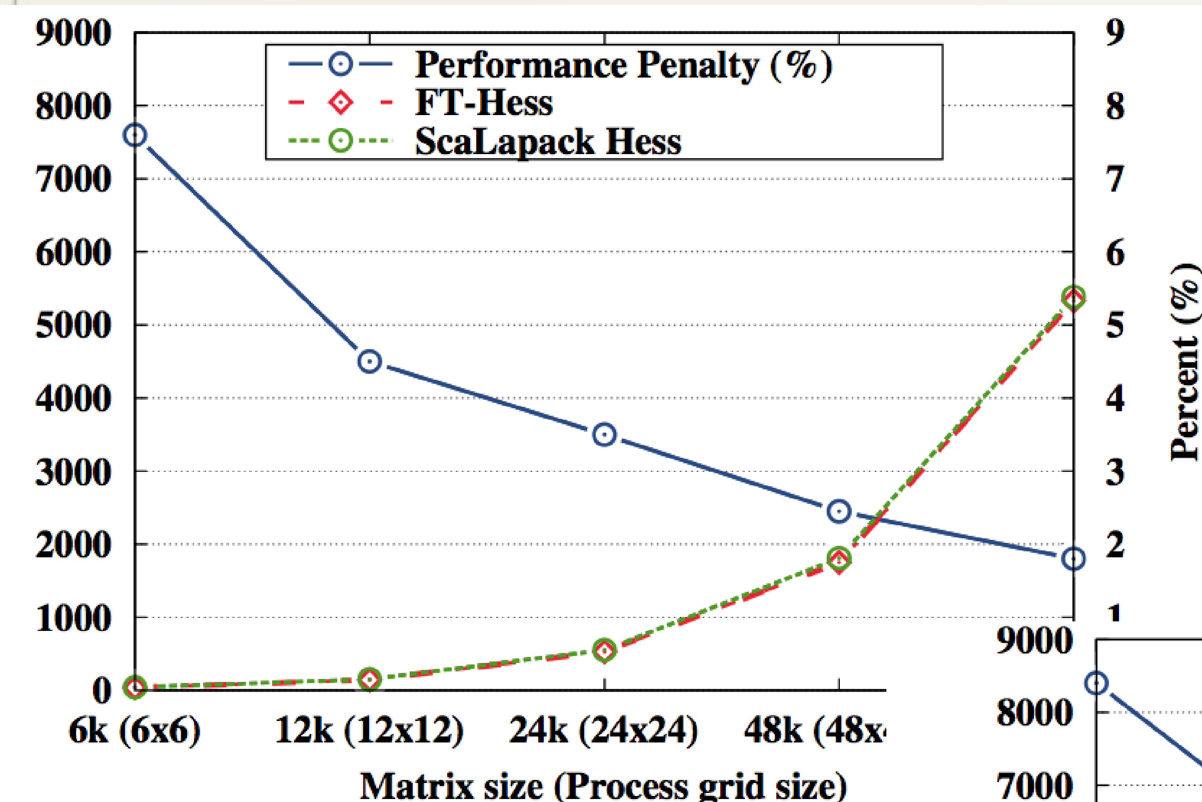


Computation for Two-Error Protection



DGETRF (solve)

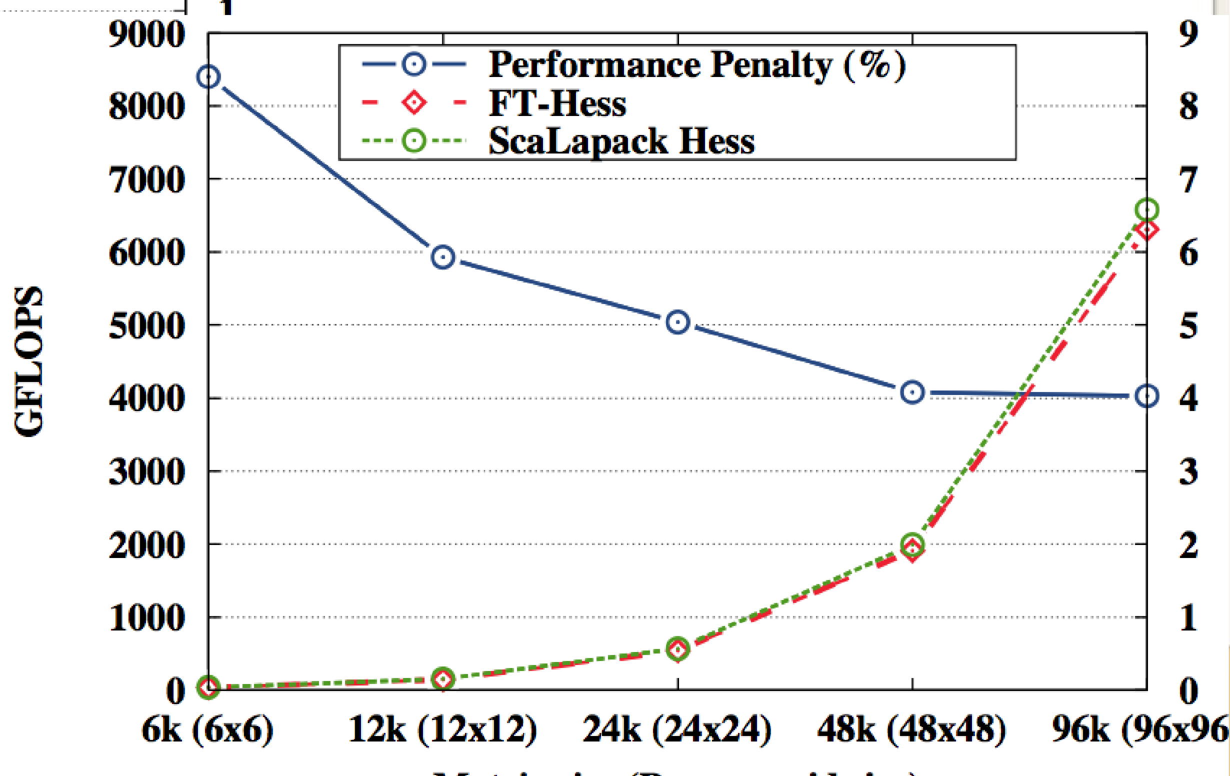
Experiment: Hessenberg Reduction



- ORNL Titan: Cray XK7
- AMD Opteron 6274
- 299,008 core
- Cray Gemini 3D Torus

No errors

1 error

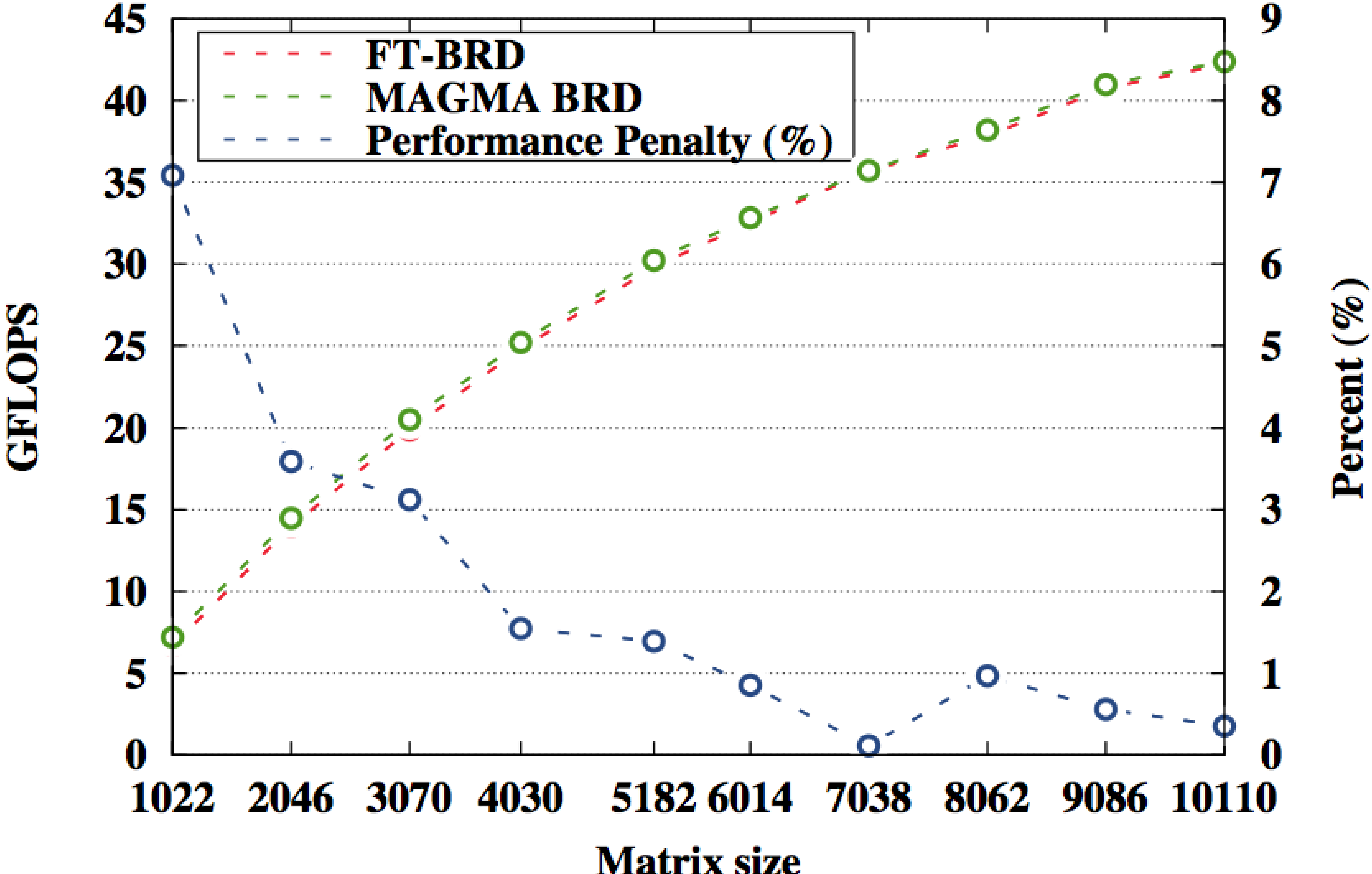


Hessenberg Reduction: Recovery Error

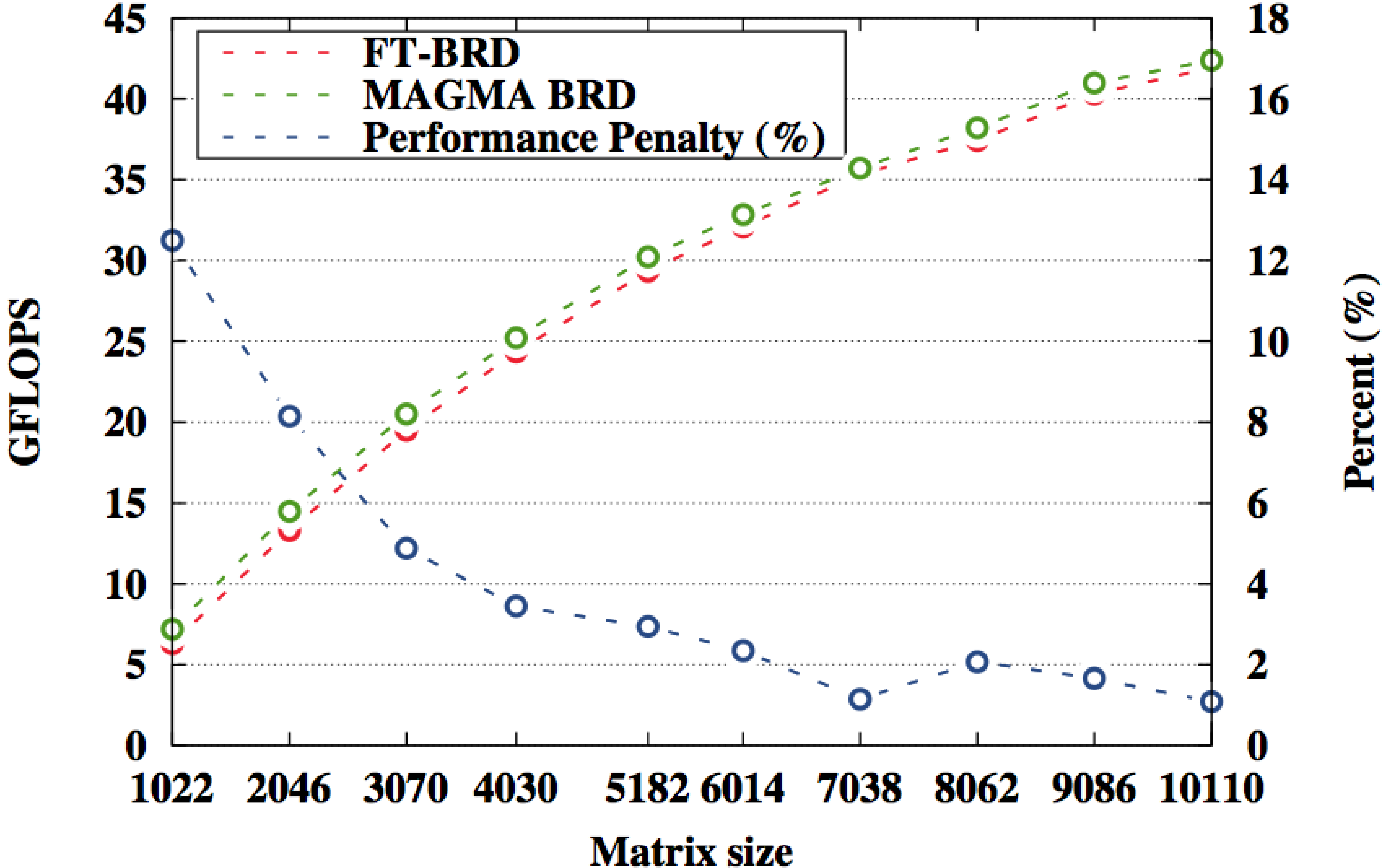
Cores	Process grid	Fault-Tolerant	Classic (ScaLAPACK)
36	6 x 6	5.20×10^{-3}	5.01×10^{-3}
144	12 x 12	3.09×10^{-3}	2.34×10^{-3}
576	24 x 24	2.16×10^{-3}	1.17×10^{-3}
2304	48 x 48	1.36×10^{-3}	6.35×10^{-4}
9216	96 x 96	1.03×10^{-3}	3.37×10^{-4}

$$r_\infty = \frac{\|A - QHQ^T\|_\infty}{\|A\|_\infty N \epsilon}$$

Experiment: BRD on GPU (no error)



Experiment: BRD on GPU (1 error)



Further Details

icl.utk.edu/ft-la

ICL UT INNOVATIVE COMPUTING LABORATORY THE UNIVERSITY of TENNESSEE

Questions

- What is the cost of error recovery: error early vs. error late?
- 50% of “our” errors are in the same column/row. What to do?
- Don't call it checksum, call it L1-norm regularization
- What is the cost of 2 errors? K errors? Is it K^3?
- How does MTBF change if the machine/software can tolerate 1 error?