Has the time come to move from petaflops (on simple benchmarks) to petadata per unit of time and power (on sophisticated benchmarks)?

Michael J. Flynn, Stanford University, USA
Oskar Mencer, Imperial College, London, UK
Veljko Milutinovic, University of Belgrade, Serbia
Per Stenstrom, Chalmers University of Technology, Sweden
Goran Rakocevic, Mathematical Institute Belgrade, Serbia
Roman Trobec, Institute Jozef Stefan, Slovenia
Mateo Valero, Barcelona Supercomputing Centre, Spain

How to assess a computing system?

THE TOP500 LIST

TEST: LINPACK

MEASURE: FLOATING POINT OPERATIONS PER SECOND

THE GRAPH500 LIST

TEST: BREADTH-FIRST SEARCH

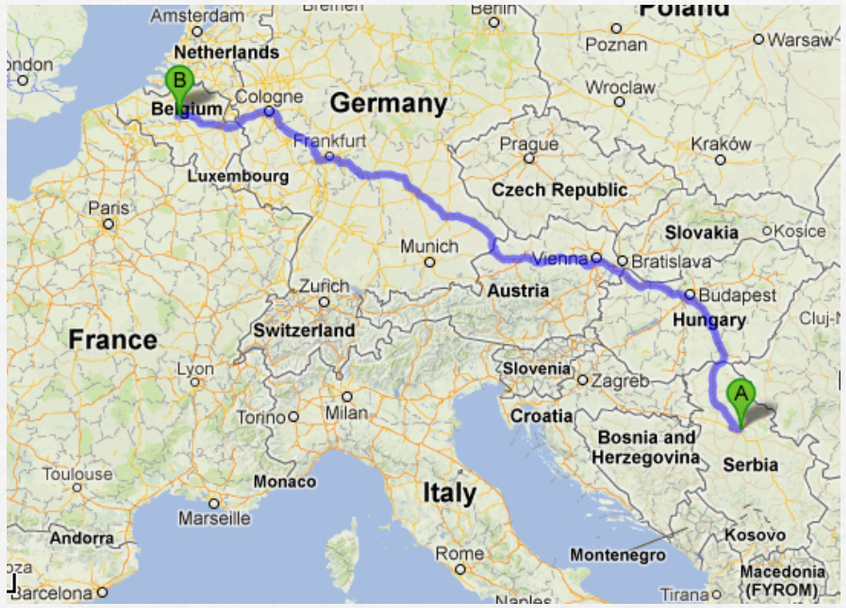
MEASURE: TRAVERSED EDGES PER SECOND

THE GREEN500 LIST

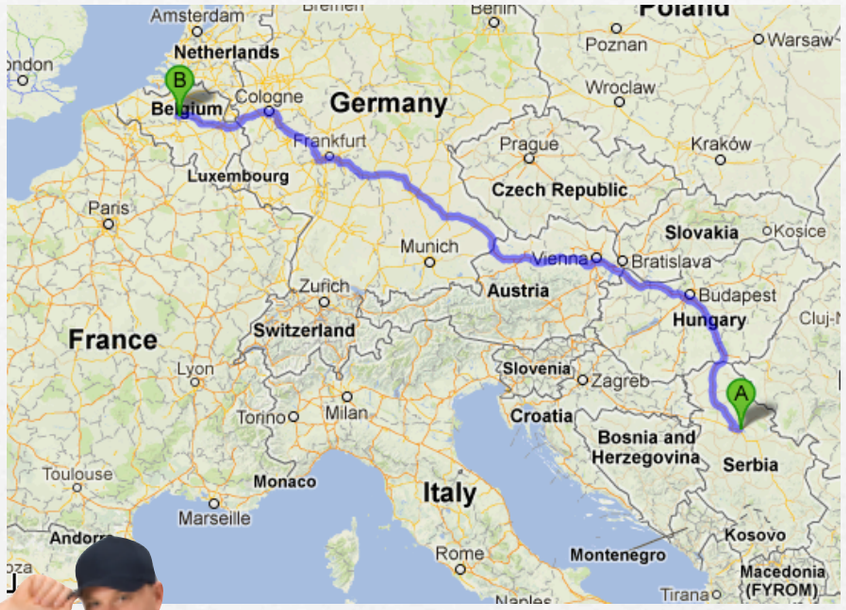
TEST: LINPACK

MEASURE: FLOATING POINT OPERATIONS PER SECOND PER WATT

Fastest ground transport?

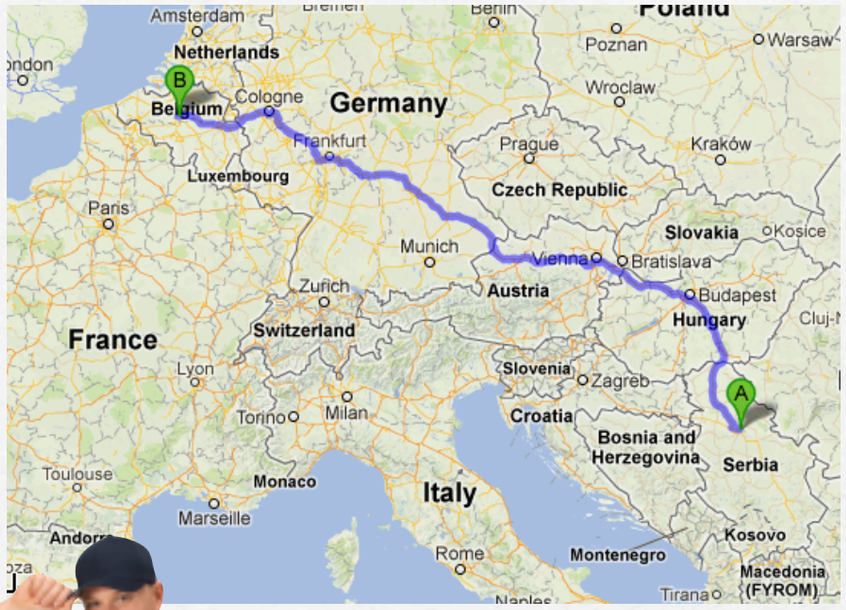


Fastest ground transport?



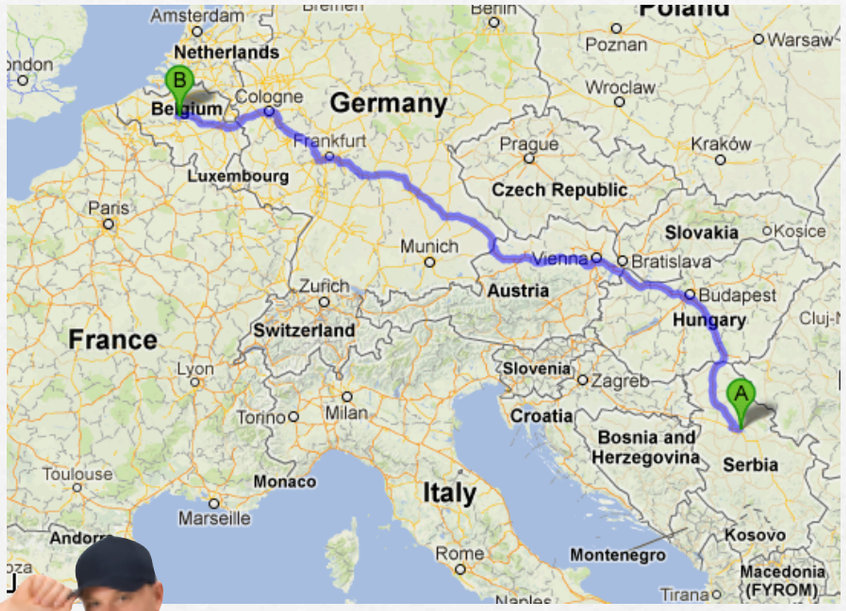
4/13

Fastest ground transport?



5/13

Fastest ground transport?



6/13

Measure multiple aspects on multiple benchmarks?

DOD HPC MODERNIZATION PROGRAM

TEST: 15 DIFFERENT CODES

MEASURE: APPLICATION RUNTIME

SUSTAINED SYSTEM PERFORMANCE

TEST: 7 DIFFERENT CODES

MEASURE: EFFECTIVE SYSTEM PERFORMANCE (ESP) METRIC

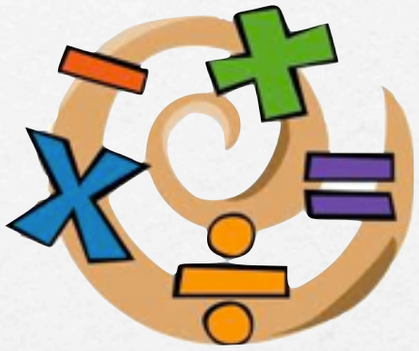
but, will a requirement port?

8/13

Several thoughts on a metric

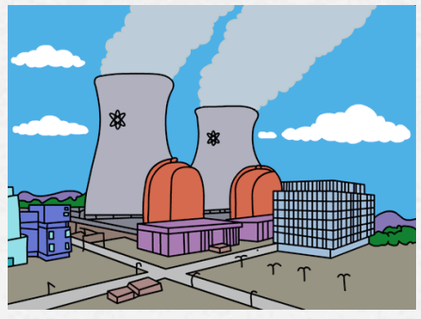
AVOID SPECIFIC OPERATIONS

- DO WE REALLY CARE HOW MANY FP OPS A SYSTEM CAN DO?
- HOW MANY WILL IT DO IN SOLVING MY PROBLEM?



CONSIDER THE POWER

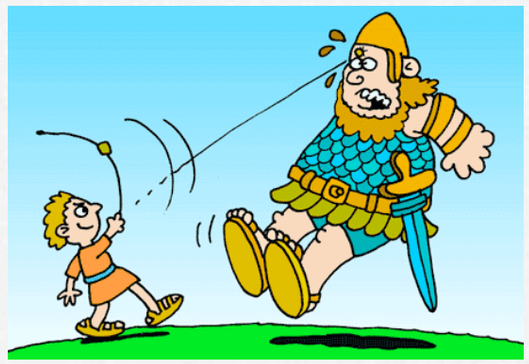
- THE SYSTEM CAN DO ALL THE OPS IT LIKES, AS LONG AS IT DOESN'T COST



Several thoughts on a metric (2)

CONSIDER THE SIZE/COST

- ACQUISITION COST, BUT THIS IS NEGOTIATED



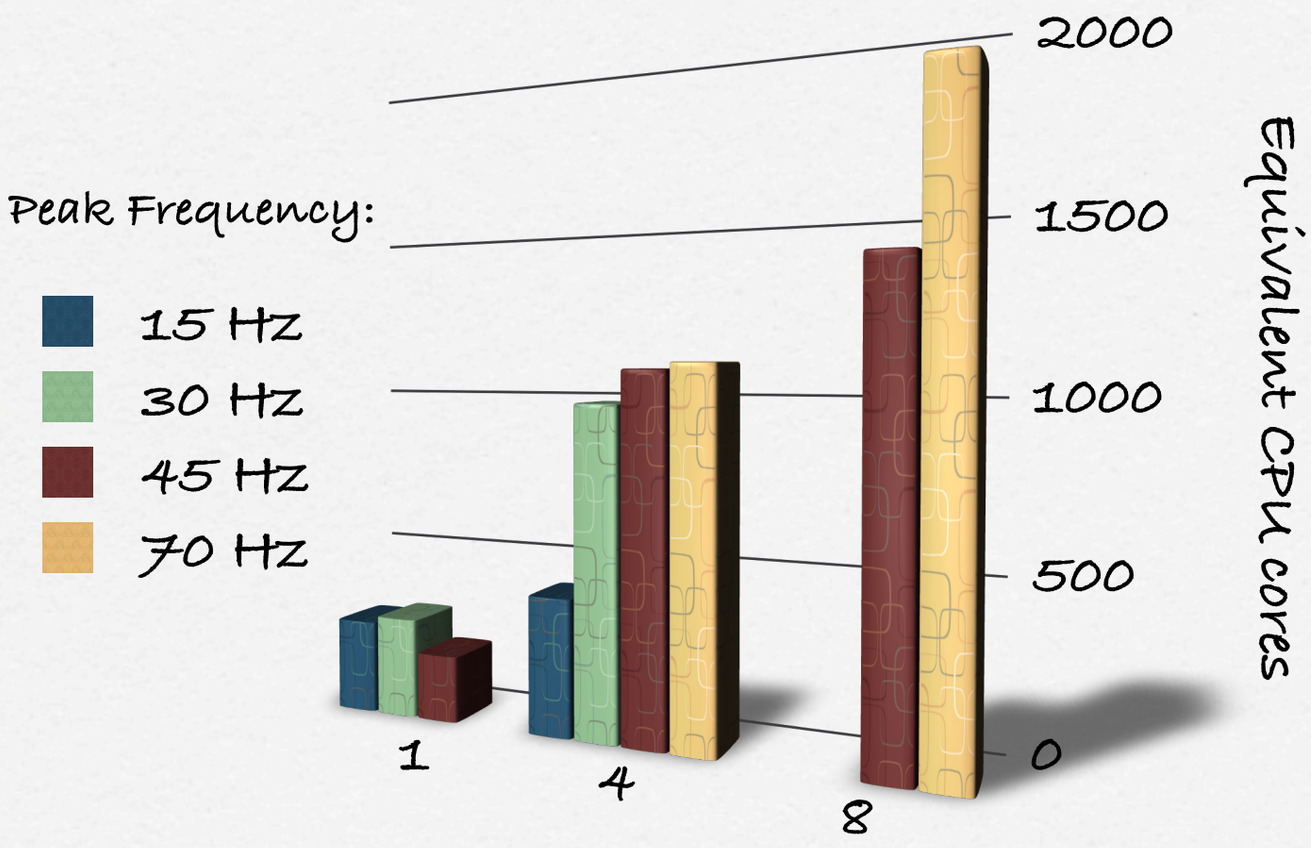
NEW BENCHMARKS?

- REAL-LIFE, FULL BLOWN APPS, NOT JUST KERNELS?

SO, WHAT DO WE COUNT?

- RUN TIME?
- DATA UNITS PER SECOND? (FOR BIG DATA PROBLEMS)

Support for our claim



Performance of Maxeler-accelerated Finite Difference Modeling

Platform	Idle	Load
Dual Xeon 2.66GHz	185W	255W
with Max 2 cards	210W	240W

Power usage published by J. P. Morgan, the credit derivatives risk calculation

Platform	Speedup
Full precision	31x
Reduced precision	37x

Speedup versus 8 core Xeon server published by J. P. Morgan, the credit derivatives risk calculation

Conclusion

- Whenever a paradigm shift happens
  in computer technology,
  computer architecture,
  or computer applications,
  a new approach has to be introduced
- The Exascale era requires this major shift!

Thank You for your attention

Communications of the ACM, May 2013