HPGMG: Relevant Benchmarking for Scientific Computing

This talk: `http://59A2.org/files/20150525-HPGMG.pdf`

Jed Brown `jed@jedbrown.org` (ANL and CU Boulder)
HPGMG Collaborators: Mark Adams, Sam Williams, John Shalf, Erich Strohmeier (LBNL)

HPCSE, Czech Republic, 2015-05-25

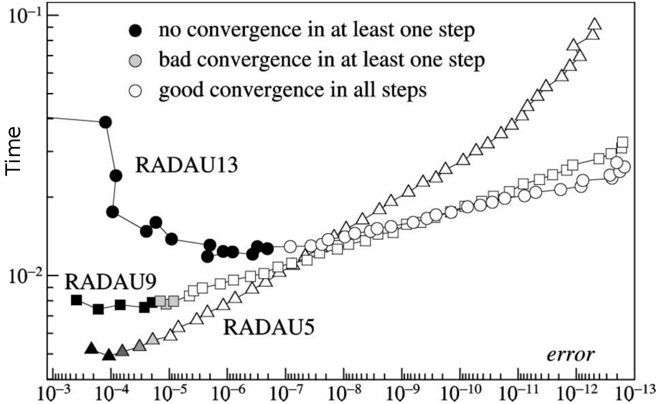
What is performance?

Dimensions

- Model complexity
- Accuracy
- Time
  - per problem instance
  - for the first instance
  - compute time versus human time
- Cost
  - incremental cost
  - subsidized?

- Terms relevant to scientist/engineer
- Compute meaningful quantities – needed to make a decision or obtain a result of scientific value—not one iteration/time step
- No flop/s, number of elements/time steps

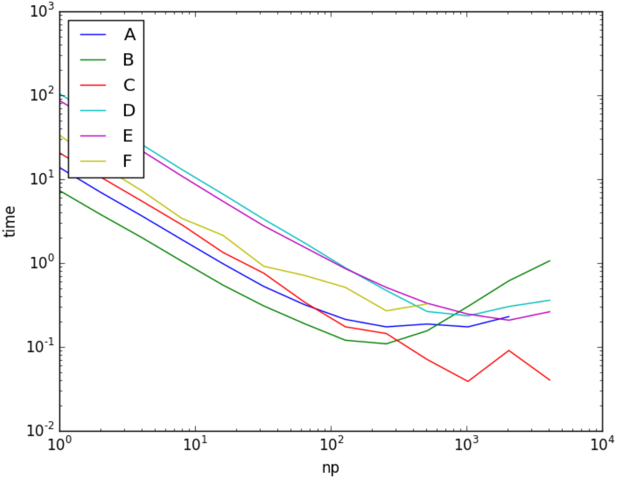
Work-precision diagram: *de rigueur* in ODE community



[Hairer and Wanner (1999)]

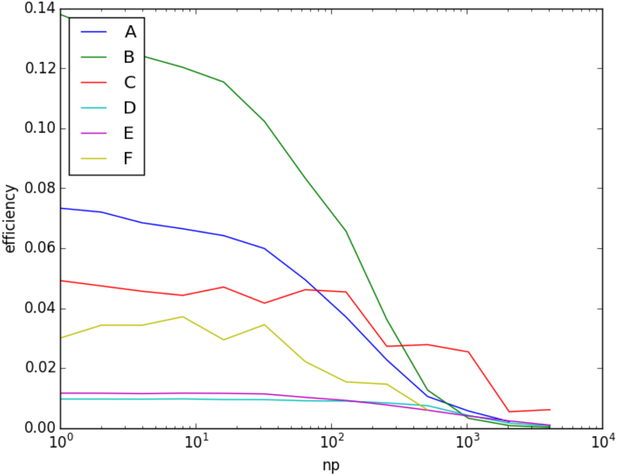
- Tests discretization, adaptivity, algebraic solvers, implementation
- No reference to number of time steps, flop/s, etc.
- Useful performance results inform *decisions* about *tradeoffs*.

Strong Scaling: cost-time tradeoff



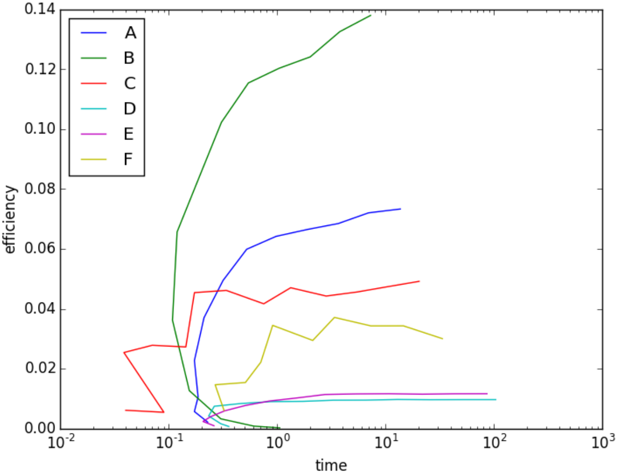
- Good: shows absolute time
- Bad: log-log plot makes it difficult to discern efficiency
  - Stunt 3: `http://blogs.fau.de/hager/archives/5835`
- Bad: plot depends on problem size

Strong Scaling: cost-time tradeoff



- Good: shows efficiency at scale
- Bad: no absolute time, depends on problem size

Strong Scaling: cost-time tradeoff



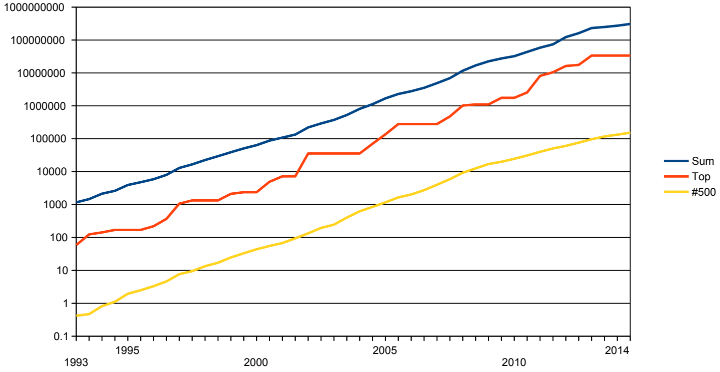
- Good: absolute time, absolute efficiency (like DOF/s/cost)
- Good: independent of problem size for perfect weak scaling
- Bad: hard to see machine size (but less important)

Exascale Science & Engineering Demands

- Model fidelity: resolution, multi-scale, coupling
  - Transient simulation is not weak scaling: $\Delta t \sim \Delta x$
- Analysis using a sequence of forward simulations
  - Inversion, data assimilation, optimization
  - Quantify uncertainty, risk-aware decisions
- Increasing relevance $\implies$ external requirements on time
  - Policy: 5 SYPD to inform IPCC
  - Weather, manufacturing, field studies, disaster response
- "weak scaling" [. . .] will increasingly give way to "strong scaling"

  [The International Exascale Software Project Roadmap, 2011]
- ACME @ 25 km scaling saturates at $< 10\%$ of Titan (CPU) or Mira
  - Cannot decrease Δx: SYPD would be too slow to calibrate
  - "results" would be meaningless for 50-100y predictions, a "stunt run"
- **ACME v1 goal of 5 SYPD is pure strong scaling.**
  - Likely faster on Edison (2013) than any DOE machine –2020
  - Many non-climate applications in same position.

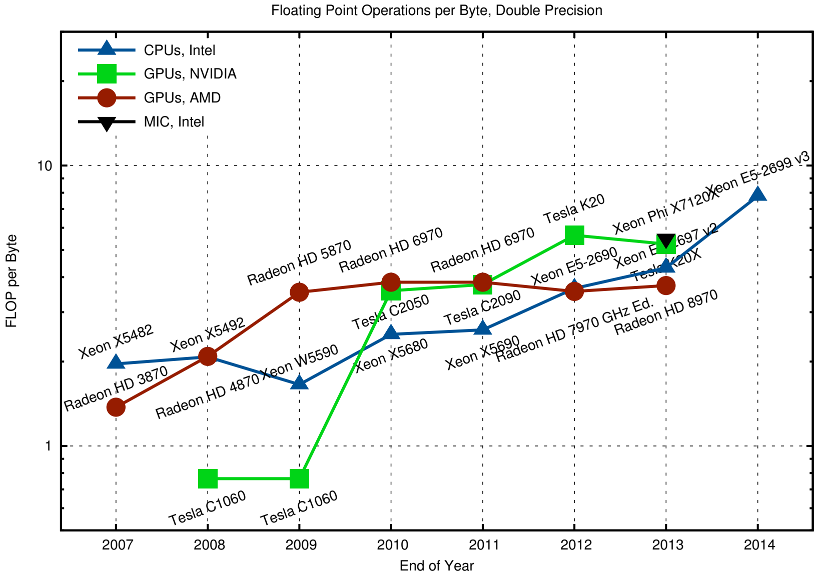
HPL and the Top500 list



- High Performance LINPACK
- Solve $n \times n$ dense linear system: $\mathscr{O}(N^{3/2})$ flops on $N = n^2$ data
- Top500 list created in 1993 by Hans Meuer, Jack Dongarra, Erich Strohmeier, Horst Simon

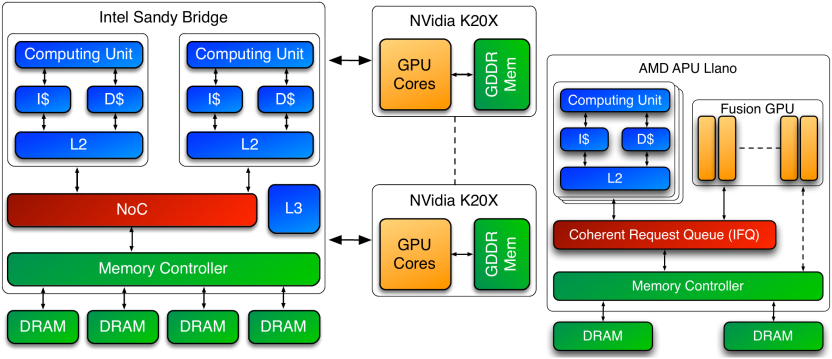
Role of HPL

- The major centers have their own benchmark suites (e.g., CORAL)
- Nobody (vendors or centers) will say they built an HPL machine
- HPL ranking and peak flop/s are still used for press releases
- Machines need to be justified to politicians holding the money
  - Politicians are vulnerable to propaganda and claims of inefficient spending
- It is naive to believe HPL has no influence on procurement or on scientists' expectations


Floating Point Operations per Byte, Double Precision
CPUs, Intel
GPUs, NVIDIA
GPUs, AMD
MIC, Intel
FLOP per Byte
10
1
Xeon X5482
Xeon X5492
Xeon W5590
Xeon X5680
Xeon X5690
Xeon E5-2690
Xeon E5-2697 v2
Xeon E5-2699 v3
Xeon Phi X7120X
Radeon HD 3870
Radeon HD 4870
Radeon HD 5870
Radeon HD 6970
Radeon HD 6970
Radeon HD 7970 GHz Ed.
Radeon HD 8970
Tesla C1060
Tesla C1060
Tesla C2050
Tesla C2090
Tesla K20
Tesla K20X
2007
2008
2009
2010
2011
2012
2013
2014
End of Year

[c/o Karl Rupp]

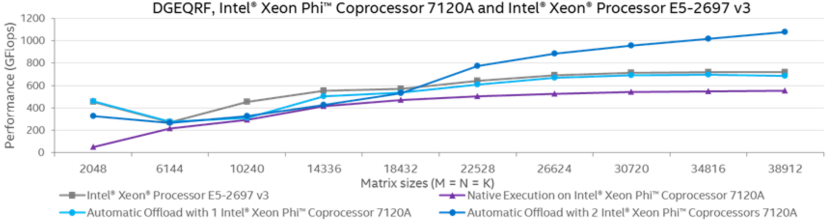
It’s all about the memory

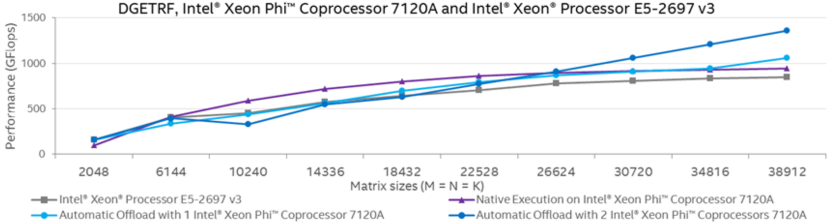


[Ang et al, 2014]

- Memory motion dominates floating point cost
- About half of die devoted to caches
- Network moving on-die, maybe throughput cores
- High-bandwidth on-package memory may have *worse* latency than DRAM

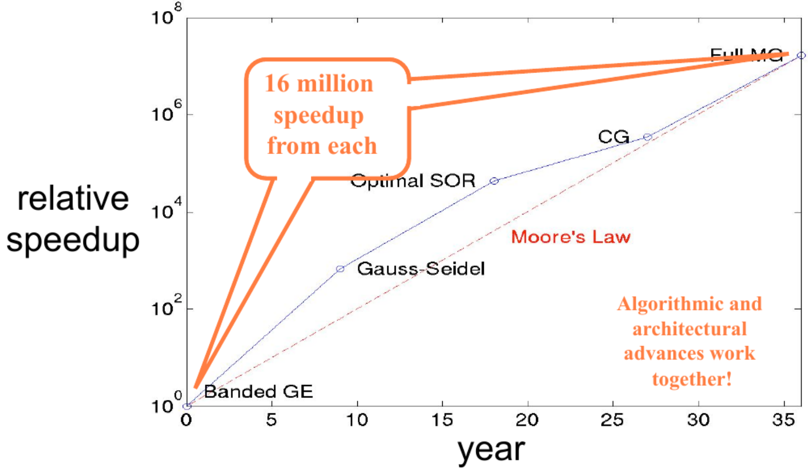
Arithmetic intensity is not enough





- QR and LU factorization have same complexity.
- Stable QR factorization involves more synchronization.
- Synchronization is much more expensive on Xeon Phi.

Algorithms keep pace with hardware (sometimes)



[c/o David Keyes]

- Opportunities now: uncertainty quantification, design
- Incentive to find optimal algorithms for more applications

What does “representative” mean?

- Diverse applications
  - Explicit PDE solvers (seismic wave propagation, turbulence)
  - Implicit PDE solvers and multigrid methods (geodynamics, structural mechanics, steady-state RANS)
  - Irregular graph algorithms (network analysis, genomics, game trees)
  - Dense linear algebra and tensors (quantum chemistry)
  - Fast methods for N-body problems (molecular dynamics, cosmology)
  - Cross-cutting: data assimilation, uncertainty quantification
- Diverse external requirements
  - Real-time, policy, manufacturing
  - Privacy
  - In-situ processing of experimental data
  - Mobile/energy limitations

Necessary and sufficient

Goodhart's Law

When a measure becomes a target, it ceases to be a good measure.

- Features stressed by benchmark **necessary** for some apps
- Performance on benchmark **sufficient** for most apps

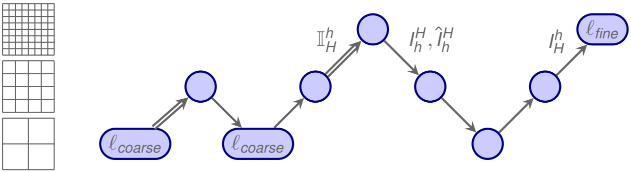
HPGMG: a new benchmarking proposal

- `https://hpgmg.org`, hpgmg-forum@hpgmg.org mailing list
- Mark Adams, Sam Williams (finite-volume), Jed (finite-element), John Shalf, Brian Van Straalen, Erich Strohmeier, Rich Vuduc
- Gathering momentum, SC14 BoF
- Implementations

  Finite Volume memory bandwidth intensive, simple data dependencies, 2nd and 4th order

  Finite Element compute- and cache-intensive, vectorizes, overlapping writes
- Full multigrid, well-defined, scale-free problem
- Matrix-free operators, Chebyshev smoothers

Full Multigrid (FMG): Prototypical Fast Algorithm

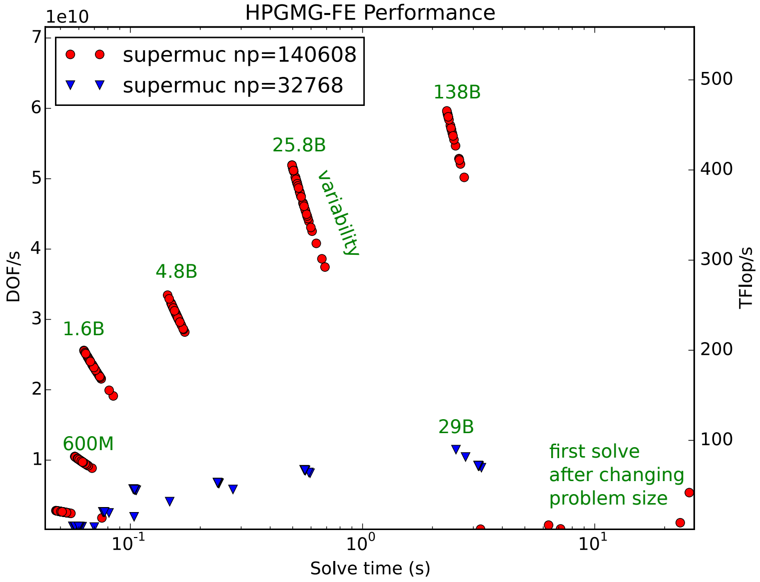


- start with coarse grid
- truncation error within one cycle
- about five work units for many problems
- no "fat" left to trim – robust to gaming
- distributed memory – restrict active process set using Z-order
  - $\mathscr{O}(\log^2 N)$ parallel complexity stresses network
- scale-free specification
  - no mathematical reward for decomposition granularity
  - don't have to adjudicate "subdomain"

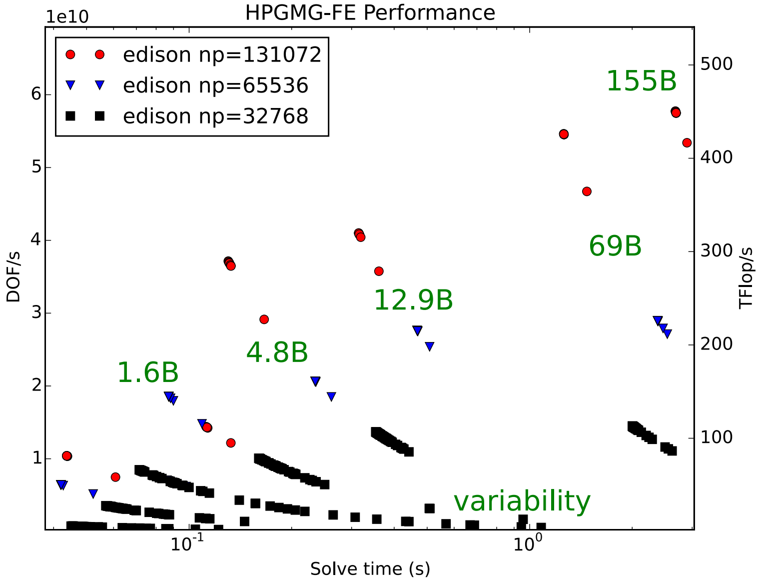
Multigrid design decisions

- Q_2 finite elements
  - Partition of work not partition of data – sharing/overlapping writes
  - Q_2 is a middle-ground between lowest order and high order
  - Matrix-free pays off, tensor-product element evaluation
- Linear elliptic equation with manufactured solution
- Mapped coordinates
  - More memory streams, increase working set, longer critical path
- No reductions
  - Coarse grid is strictly more difficult than reduction
  - Not needed because FMG is a direct method
- Chebyshev/Jacobi smoothers, $V(3,1)$ cycle
  - Multiplicative smoothers hard to verify in parallel
  - Avoid intermediate scales (like Block Jacobi/Gauss-Seidel)
- Full Approximation Scheme

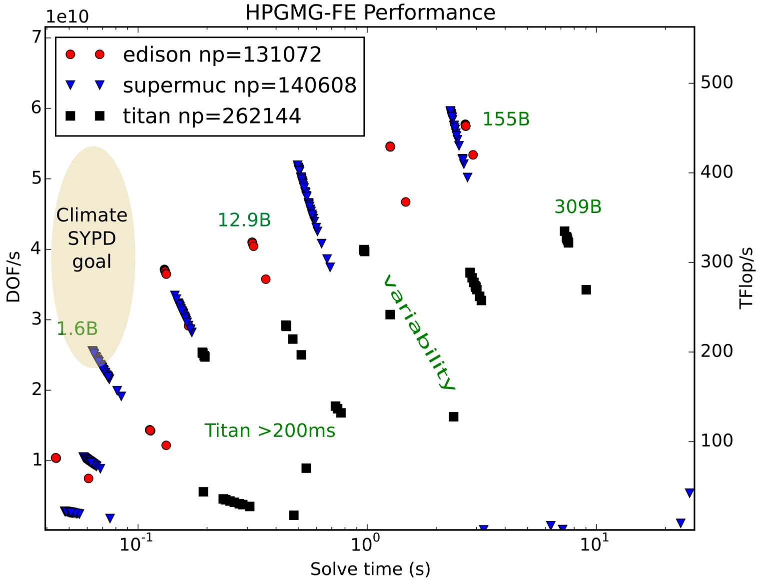
SuperMUC (FDR 10, E5-2680)



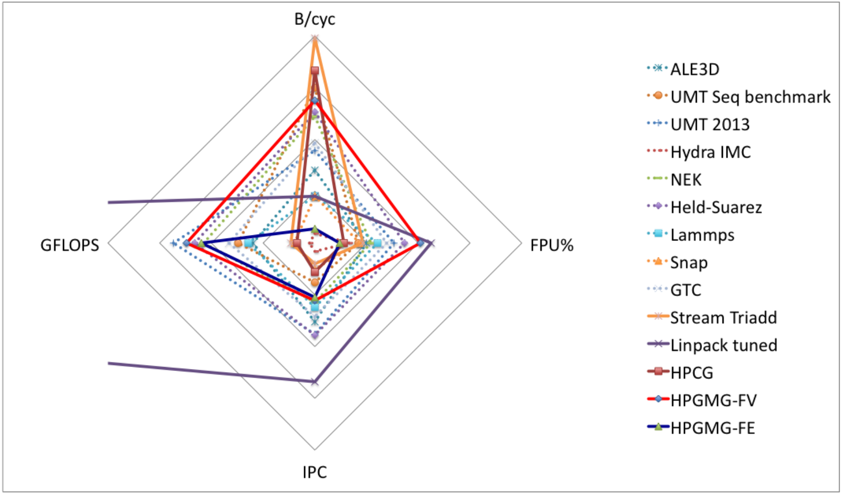
Edison (Aries, E5-2695v2)



HPGMG-FE on Edison, SuperMUC, Titan

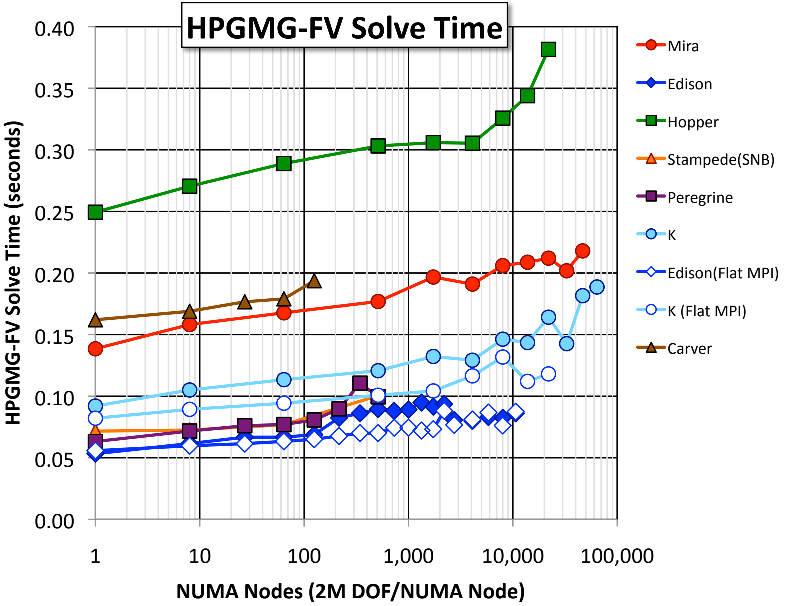


Kiviat diagrams

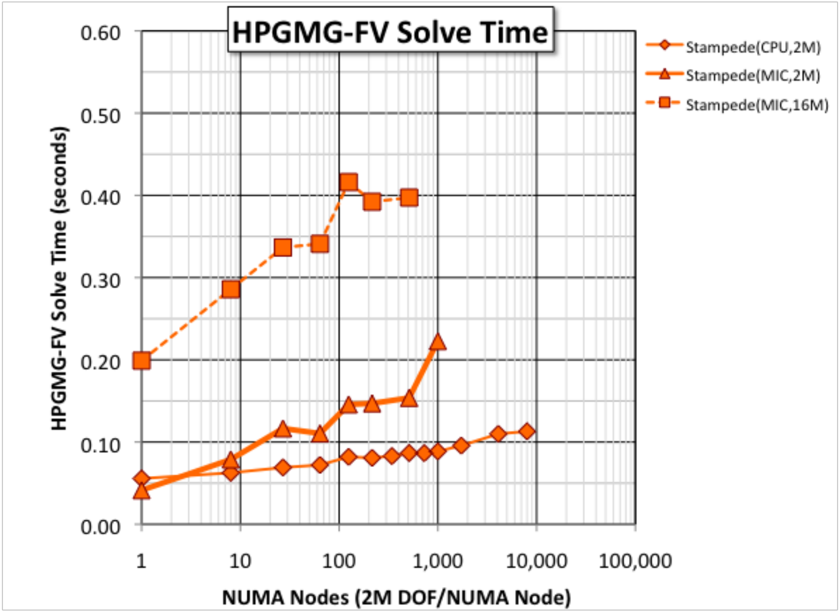


- c/o Ian Karlin and Bert Still (LLNL)

HPGMG-FV distinguishes networks at 2M DOFs/node



MIC communication bottlenecks on Stampede



Outlook

- What is the cost of performance variability?
  - Measure best performance, average, median, 10th percentile?
  - Applications bundling due to perverse queue incentives
- Should dynamic range enter into a ranking metric?
  - Why is NERSC installing DRAM in Cori?
  - Versatility is an essential part of Performance.
- Finite element or finite volume?
  - overlapping writes, cache reuse
  - FE: $> 20\%$ Intel, 6% Blue Gene/Q; vs 10% for FV
  - FV: 4th order (higher AI) improves flop/s on Intel, not on BG/Q
  - FV 4th order performs best with “red-black GS” – weak order dependence
- Linear or nonlinear?
- Irregularity and adaptivity?
- Tensor-valued coefficients?
- Elasticity?
- HPGMG does not seek to address I/O.