Lattice Field Theory

Strong Dynamics in Standard Model and Beyond

Exascale Requirements Reviews: High Energy Physics

Rich Brower (SciDAC software co-director/CUDA fellow)

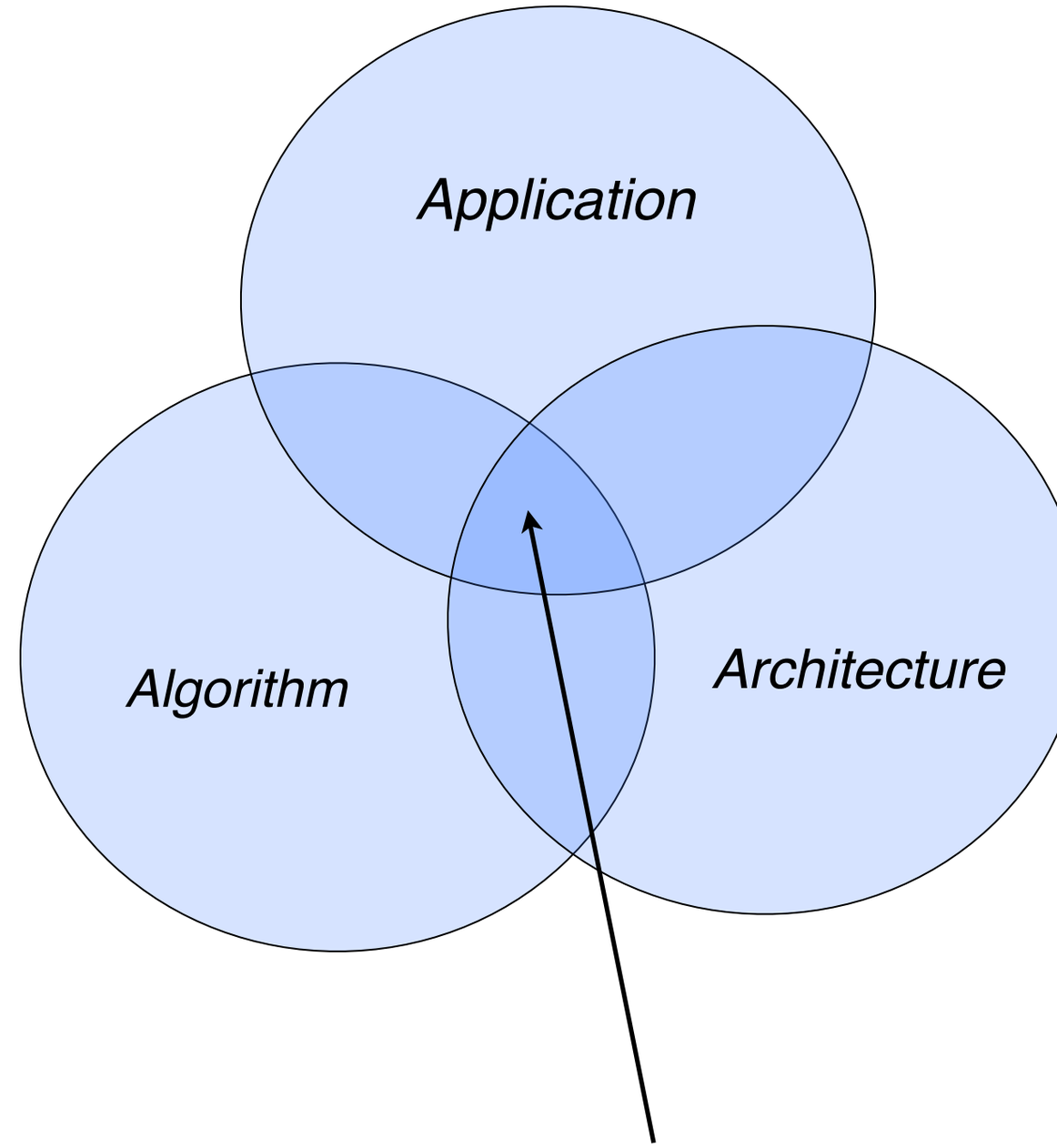
June 10, 2015

3 Part USQCD Program

- PRECISION PHYSICS

- Multi-scale ALGORITHMS

- Parallel SOFTWARE/HARDWARE

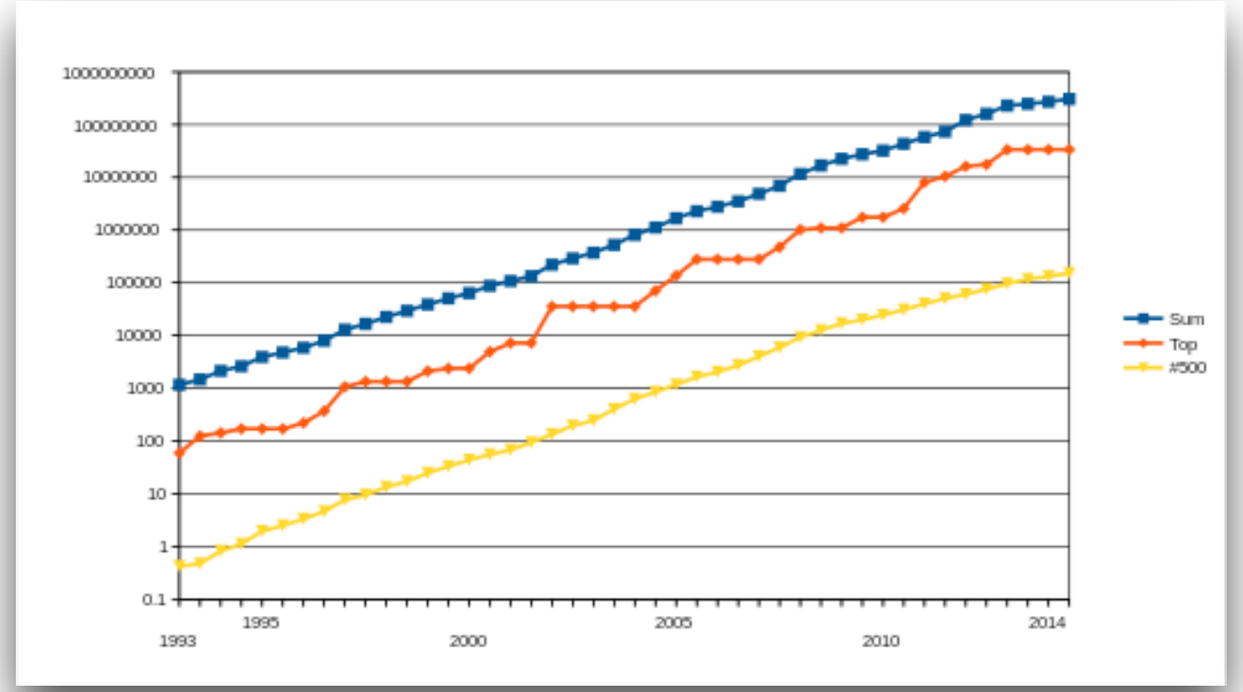


Hard to find Sweet Spot

Lattice Field Theory has Come of Age

K. Wilson: "Lecture at Lattice 1989 Capri"

"lattice gauge theory could also require a 10^8 increase in computer power AND spectacular algorithmic advances before useful interactions with experiment ...



CM-2 100 Mflops (1989) 10^7 increase in 25 years *BF/Q 1 Pflops (2012)*

Future GPU/PHI architectures will soon get us there!

What about spectacular Algorithms/Software?

HEP Need For Precision

- Fundamental advances can requiring precision theory
  - Special Relativity: Speed of light (1905)
  - General Relativity: Perihelion of Mercury (1919)
  - Quantum Field Theory: Lamb Shift (1947)
- Hadronic physics to explore Beyond the Standard Model.
  - alpha(Mz) = 0.1184 +/- 0.0007 (lattice has smallest error)
  - CKM matrix elements (several theory errors approaching expt'l)
  - g-2 seeks 0.12ppm error (Lattice uncertainty is huge challenge)
  - quark masses
  - mu2e and mu2gamma , neutrino scattering,
  - Dark matter detection through Higgs to Nucleus vertex
  - Composite Higgs and Dark Matter in BSM strong dynamics (LSD)

Lagrangian for QCD

What so difficult about this!

$$S = \int d^4x \mathcal{L}$$

$$\mathcal{L}(x) = \frac{1}{4g^2} F^{ab}_{\mu\nu} F^{ab}_{\mu\nu} + \bar{\psi}_a \delta^{ab} \gamma_\mu (\partial_\nu + A^{ab}_\mu) \psi_b + m\bar{\psi}\psi$$

- 3x3 "Maxwell" matrix field & 2+ Dirac quarks
- 1 "color" charge g & "small" quark masses m.
- Sample quantum "probability" of gluonic "plasma":

$$\text{Prob} \sim \int \mathcal{D}A_\mu(x) \det[D^\dagger_{quark}(A) D_{quark}(A)] e^{-\int d^4x F^2/2g^2}$$

"A little knowledge is a dangerous thing"

Quantum Field Theory is NOT just another math exercise solving PDEs!

	Lagrangian (i.e. PDE's)	Lattice (i.e.Computer)	Quantum Theory (i.e.Nature)
Rotational(Lorentz) Invariance	✔	✘	✔
Gauge Invariance	✔	✔	✔
Scale Invariance	✔	✘	✘*
Chiral Invariance	✔	✔	✘**

Result: QM spontaneously brakes symmetries causing (unexpected) large scales.

* color charge g is NOT a parameter! ** Small additional mass quarks (i.e. Higgs)

Algorithms to exploit multiple scales

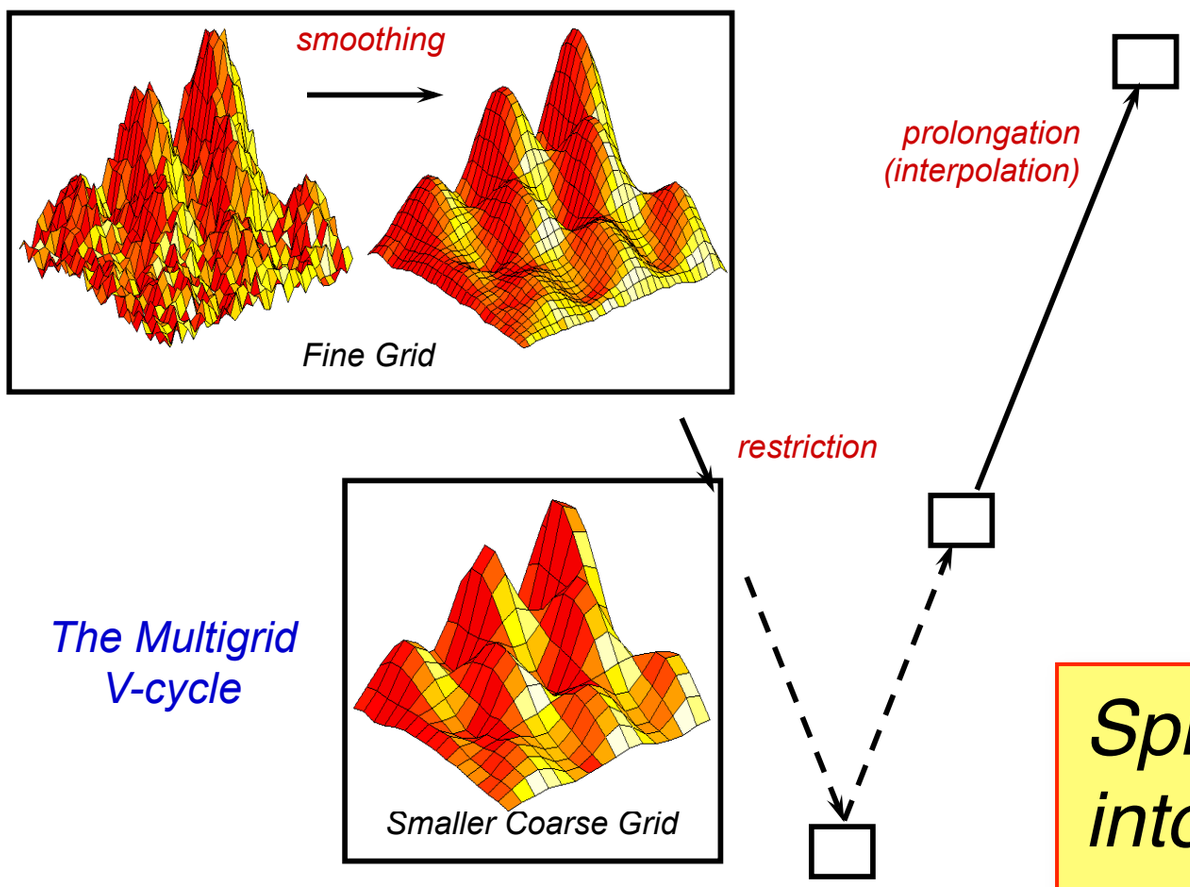
- Multigrid Linear Solvers for 3 Lattice Dirac Actions.
  - Wilson Clover
  - Domain Wall or overlap (full chirality)*
  - Staggered (partial chirality or SUSY)*
- Deflation Solvers to reduce for noise suppression
- Hybrid Monte Carlo (HMC) Evolution
  - Multi-time step Symplectic Integration
  - Rational Forces Decomposition (RHMC(
- Domain Decomposition to Communication Reduction
- ETC. (I believe the Algorithmic revolution has just begun)

* PS: Just arrived from Aspen Workshop on

"Understanding Strongly Coupled Systems in High Energy and Condensed Matter Physics"

Staggered and Domain Wall are ubiquitous on Strongly Correlated Electronic Materials!

The Multigrid Solver (at last)

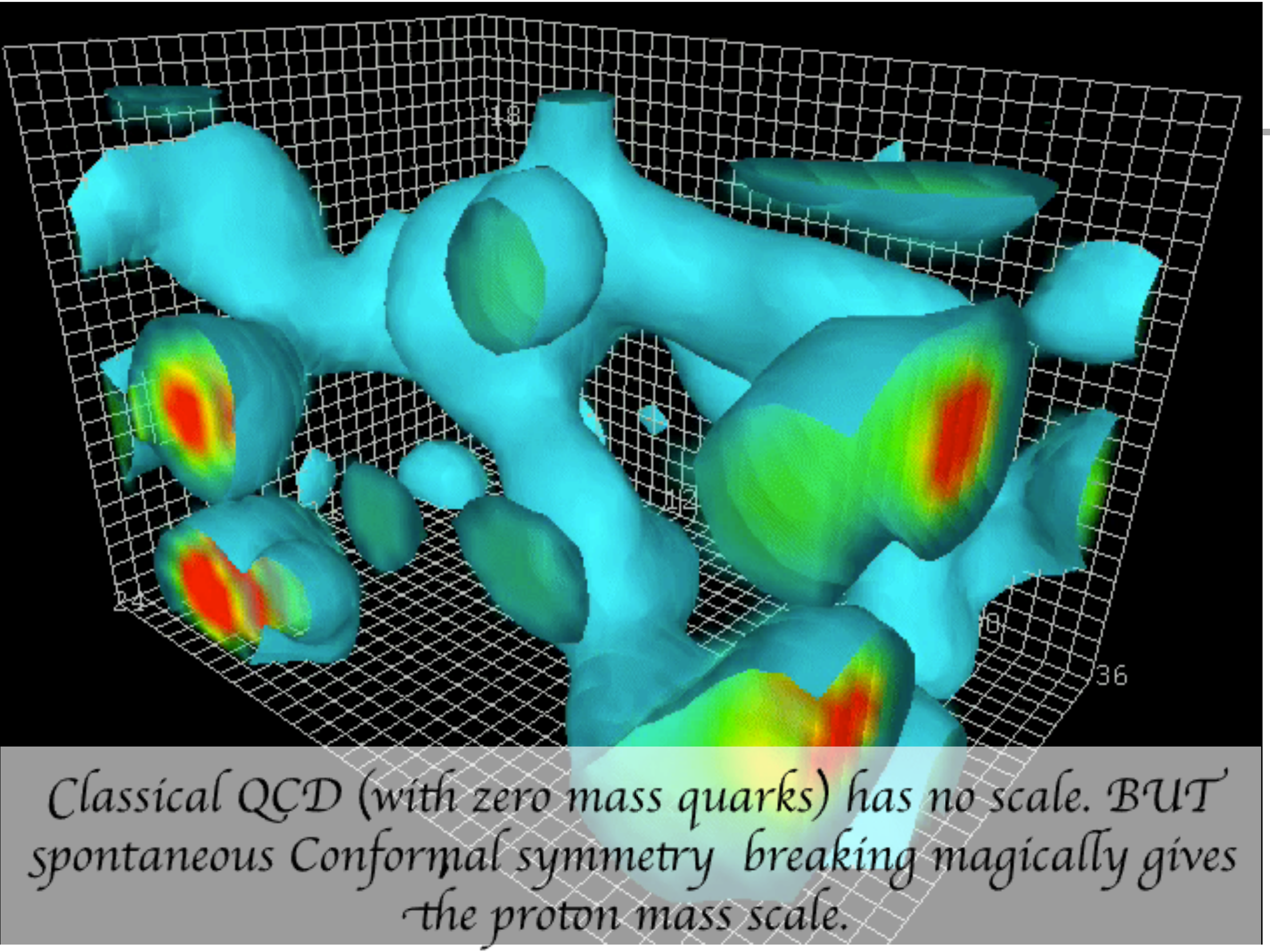


- **20 Years of QCD MULTIGRID**
  - In 1991 **Projective MG** for algorithm to long distances
  - In 2011 **Adaptive SA MG** successfully extended

Spilt the vector space into near null space $\mathcal{S}$ and the complement $\mathcal{S}_\perp$

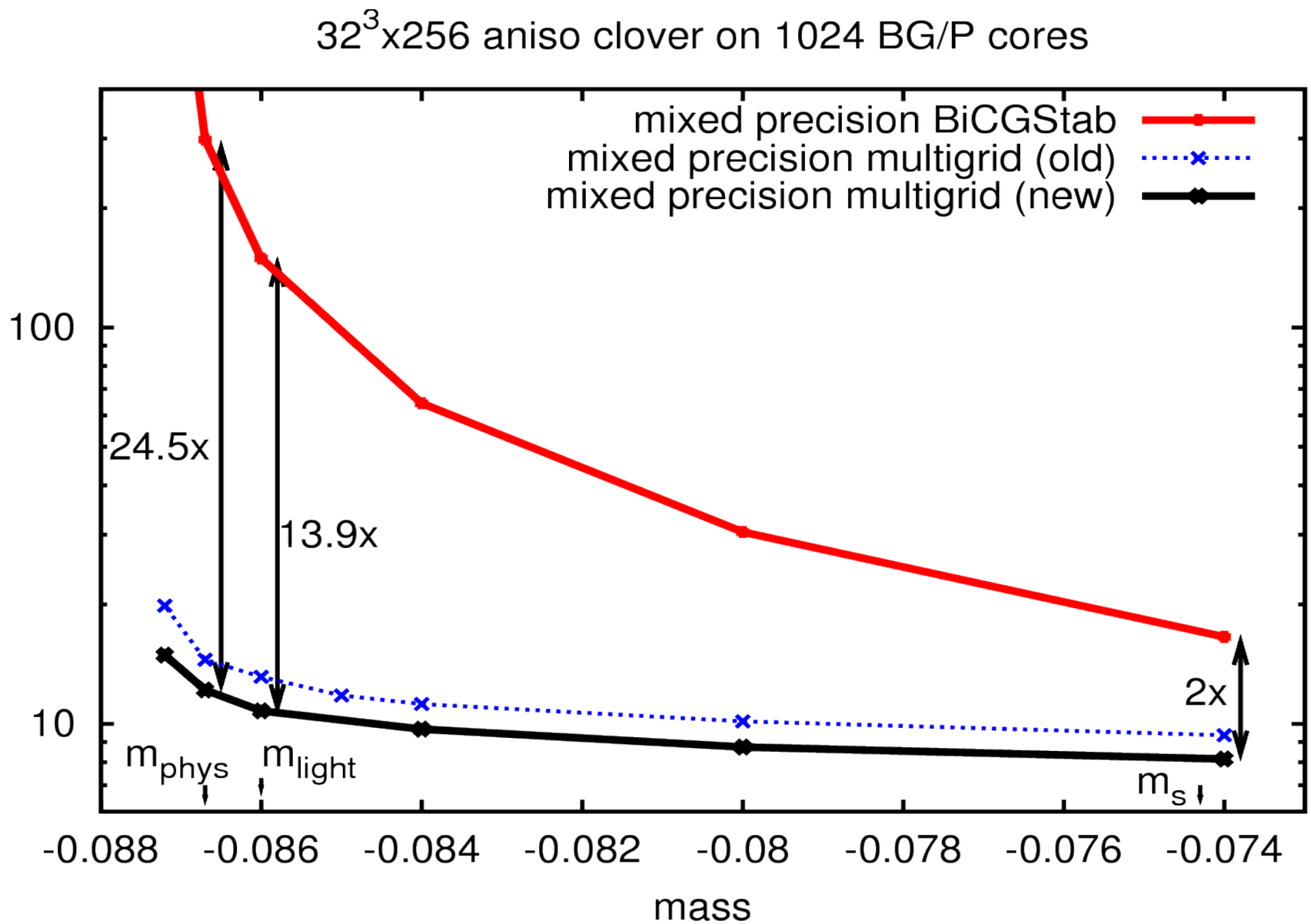
$D: \mathcal{S} \simeq 0$

Slow convergence of Dirac solver is due to small eigenvalues for vectors in near null subspace: $\mathcal{S}$.



Classical QCD (with zero mass quarks) has no scale. BUT spontaneous Conformal symmetry breaking magically gives the proton mass scale.

Multi-grid at last!
(Wilsonian Renormalization Group for Solvers)

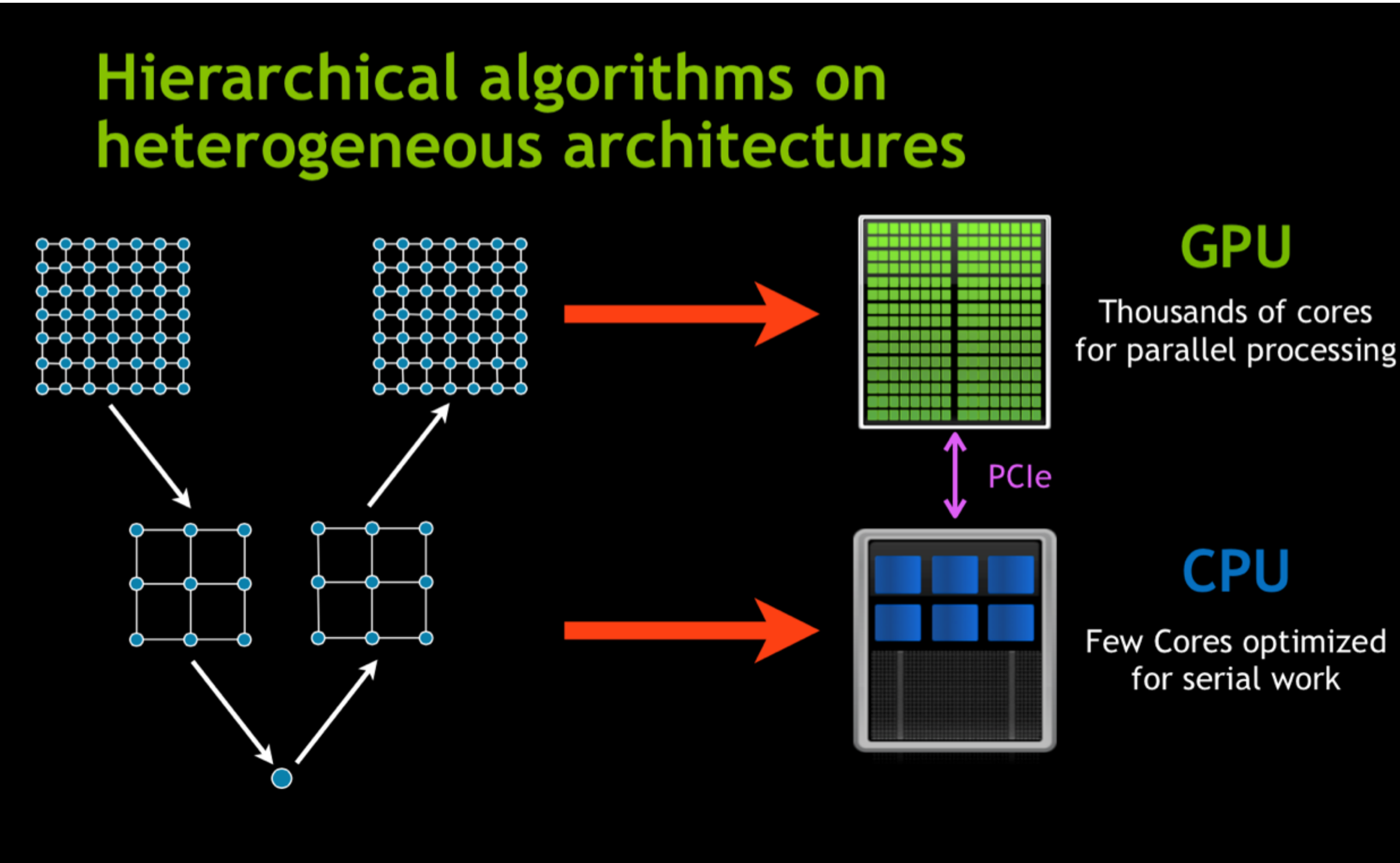


Adaptive Smooth Aggregation Algebraic Multigrid

Performance on BG/Q

"Adaptive multigrid algorithm for the lattice Wilson-Dirac operator" R. Babich, J. Brannick, R. C. Brower, M. A. Clark, T. Manteuffel, S. McCormick, J. C. Osborn, and C. Rebbi, PRL. (2010).

Mapping Multi-scale Algorithms to Multi-scale Architecture

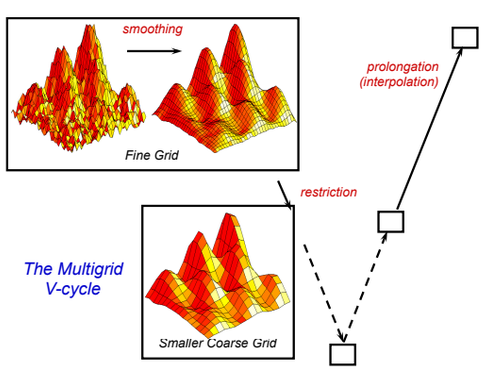


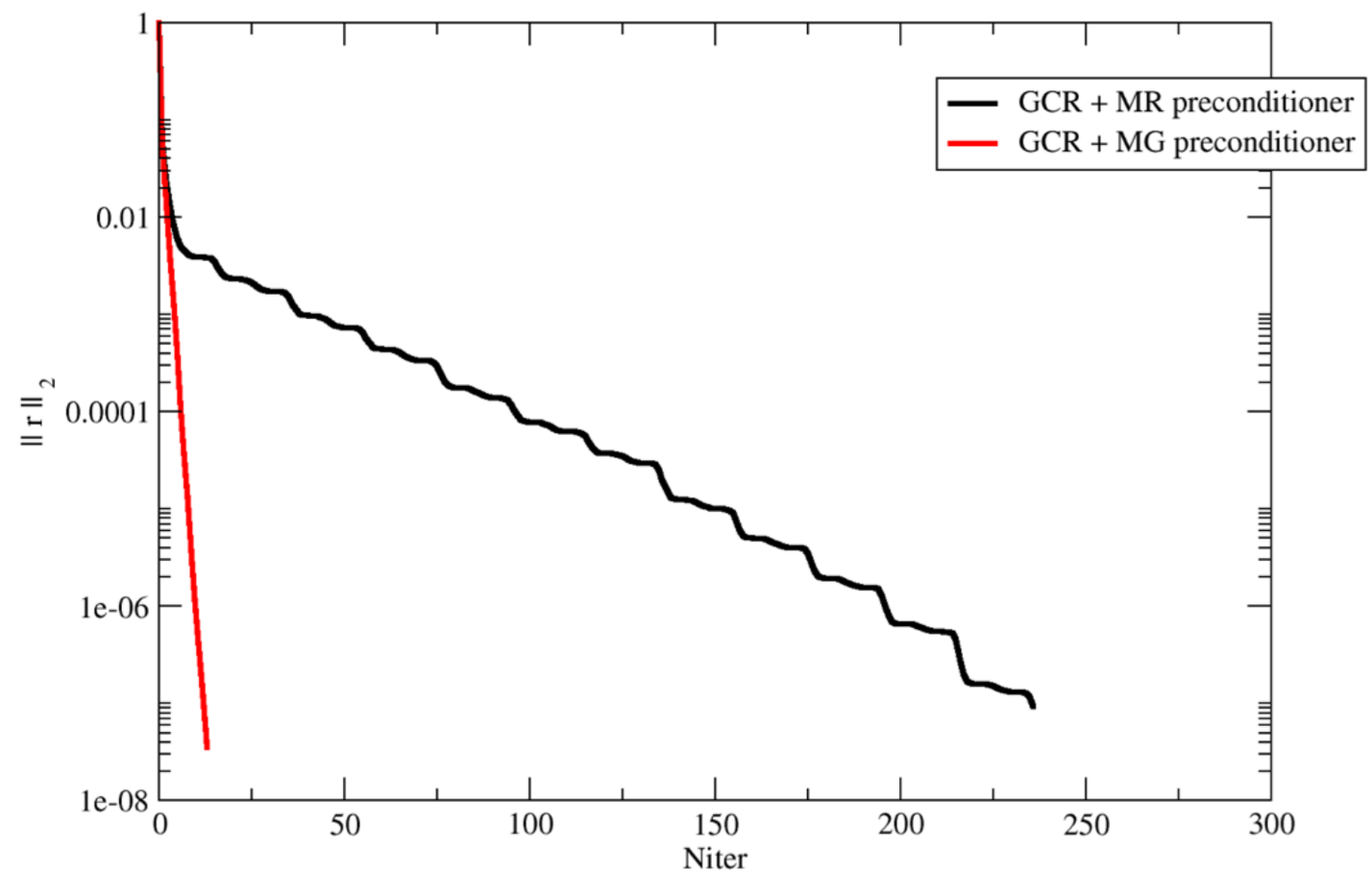
Multigrid on multi-GPU (then Phi):

Problem: Wilson MG for Light Quark beats QUDA CG solver GPUs!

Solution: Must put MG on GPU of course

+ =>



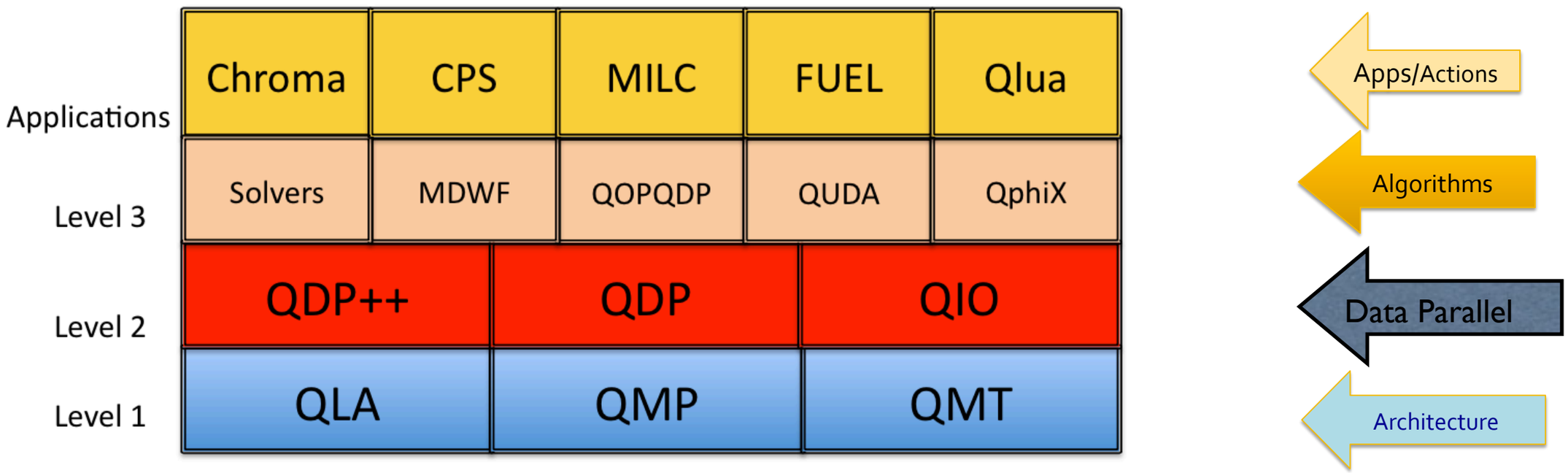


$ per solve reduced by more than 100X

GPU + MG will reduce $ cost by O(100) : see Rich Brower Michael Cheng and Mike Clark, Lattice 2014

USQCD Software Stack

On line distribution: http://usqcd.jlab.org/usqcd-software/



Chroma = 4856 files

Wilson clover

CPS = 1749 files

Domain Wall

MILC = 2300 files

Staggered

QUDA/python = 221 files

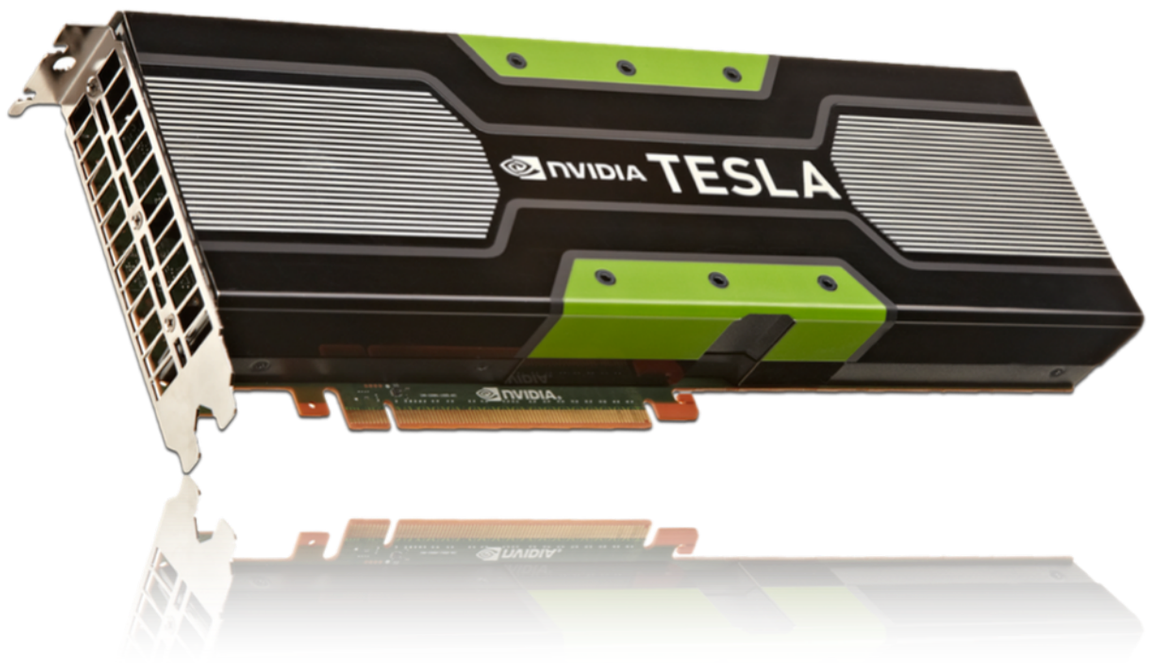
QLA/perl = 23000 files

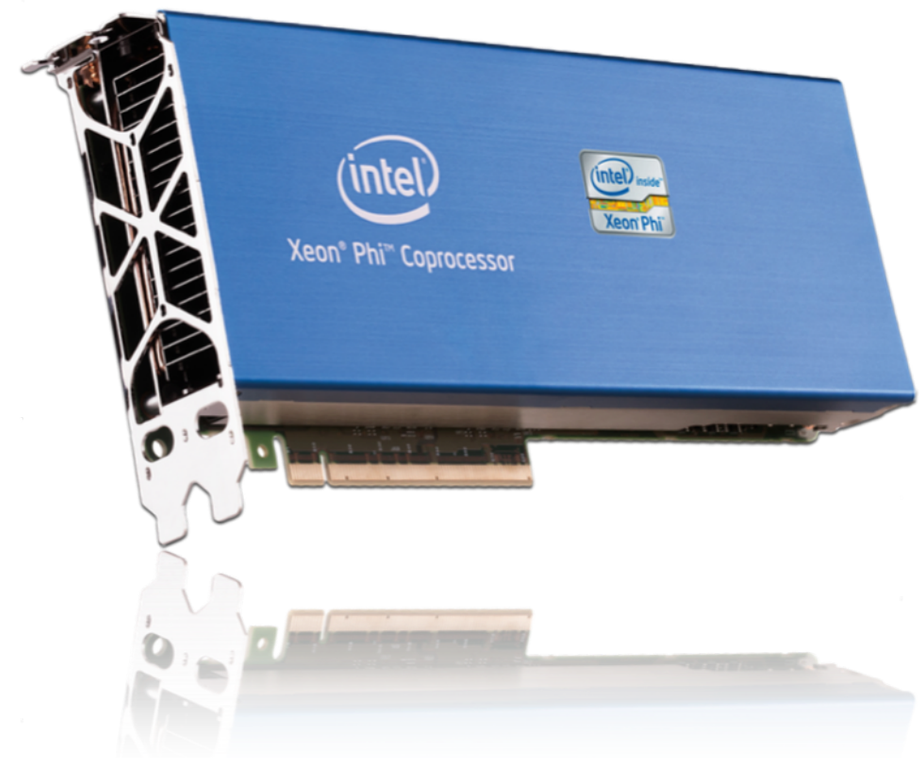
Current Highest Priority is moving to new architecture!

Applications	Chroma	CPS	MILC	FUEL	Qlua
Level 3	Solvers	MDWF	QOPQDP	QUDA	QphiX
Level 2	QDP++		QDP	QIO	
Level 1	QLA		QMP	QMT	

(May you live in Interesting Times!)

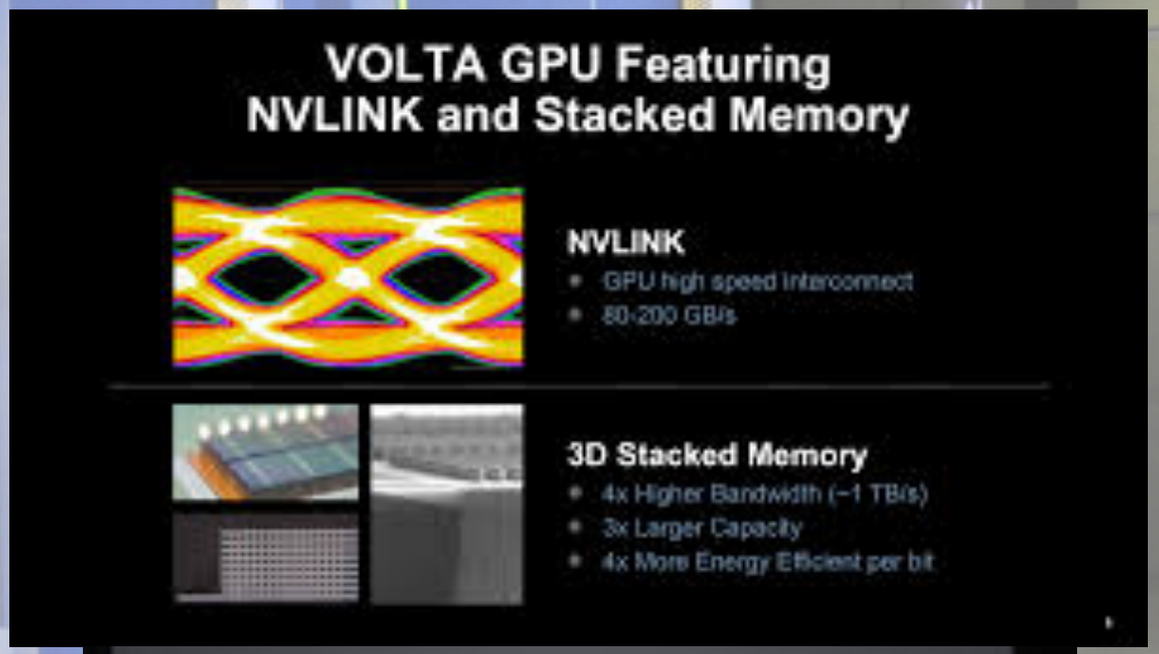
Multi-core Revolution.

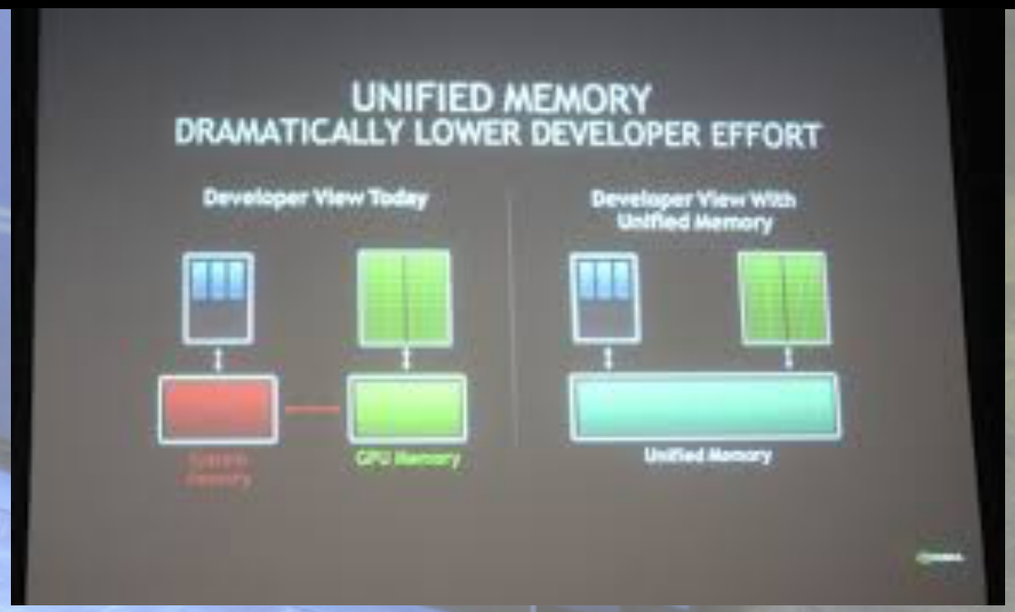




- The CORAL initiative in next two years will coincide with both NVIDIA/IBM and INTEL/CRAY rapidly evolving their architectures and programming environment with unified memory, higher bandwidth to memory and interconnect etc.
- USQCD has Strong Industrial Collaborations with NVIDIA, INTEL and IMB: former Lattice Gauge Theorist and direct access to industry engineering and software professionals.

Near FUTURE: CORAL





Algorithmic Challenges in the Multi-scale Era

One: Keeping pace with current platforms!

BG/Q(IBM) ,Titan (NVIDIA), Stampede (Intel),CORAL, …

Two: Exposing and Exploiting Multi-scale Physics

Nucleon Structure, Hadronic Excitations, Hot QCD, Nuclei etc.

Three: Conforming Physics to Hardware.

"Physics" and "Architecture" is multi-scaled but not trivially conformat.

Exascale Future Challenge

- Clearly Algorithms will advance tremendously
  - Expect the Unexpected: progress in Quantum FEM, Fermionic sign problem?
- THE Challenge is increasingly complex Software
  - Specialized Node specific Libraries QUDA/QphiX) will persist BUT there must be generic solutions as well for the majority of the Code Base.
- New (Data Parallel) DSP frameworks are being actively Explored:
  - FUEL (Lua Framework Argonne/James Osborn),
  - Xgrid (LBL/ Edinburg/Peter Boyle/Chulwoo Jung)
  - MPI/OpenMP4 GPU-PHI combined infrastructure.
  - Restructure Chroma(JLab), MILC in QDPDOP (Argonne/FermiLab)
- Of course Data Parallel Compiler and Libraries “should” be delivered with the Hardware!
  - like CMSSL (Connection Machine Scalable Scientific Library) on ye olde Thinking Machine! Why not?

Major USQCD Software contributors 2012-15

- ANL: James Osborn, Meifeng Lin, Heechang Na
- BNL: Frithjof Karsch, Chulwoo Jung, Hyung-Jin Kim,S. Syritsyn,Yu Maezawa
- Columbia: Robert Mawhinney, Hantao Yin
- FNAL: James Simone, Alexei Strelchenko, Don Holmgren, Paul Mackenzie
- JLab: Robert Edwards, Balint Joo, Jie Chen, Frank Winter, David Richards
- W&M/UNC: Kostas Orginos, Andreas Stathopoulos, Rob Fowler (SUPER)
- LLNL: Pavlos Vranas, Chris Schroeder, Rob Faulgot (FASTMath), Ron Soltz
- NVIDIA: Mike Clark, Ron Babich, Mathias Wagner
- Arizona: Doug Toussaint, Alexei Bazavov
- Utah: Carleton DeTar, Justin Foley
- BU: Richard Brower, Michael Cheng, Oliver Witzel
- MIT: Pochinsky Andrew, John Negele,
- Syracuse: Simon Catterall, David Schaich
- Washington: Martin Savage, Emanuell Chang
- Many Others: Peter Boyle, Steve Gottlieb, George Fleming et al
- "Team of Rivals" (Many others in USQCD and Int'l Community volunteer to help!)

EXTRA BACKGROUND SLIDES

Lattice QCD Approach to the *ab initio* Solution

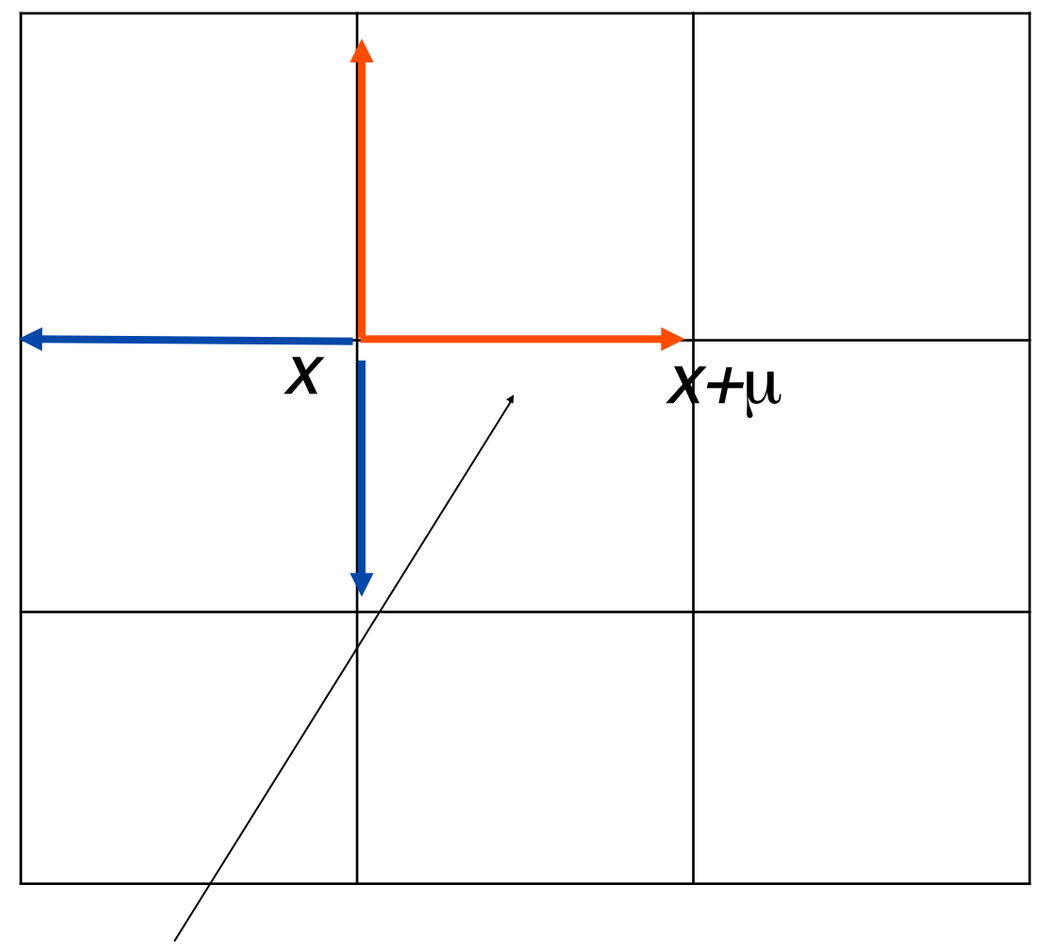
$$\mathrm{Prob}[U,\phi] = Z^{-1} e^{\beta Tr[U_{glue} + U^{\dagger}_{glue}] + \bar{\phi}(D^{\dagger}_{quark} D_{quark})^{-1}\phi}$$

$$U_{glue}(x, x+\mu) = e^{iaA_{\mu}(x)}$$

- Get rid of Determinant with "pseudo-fermions"

- Hybrid Monte Carlo (HMC): Introduces 5th "time" molecular dynamics Hamiltonian evolution.

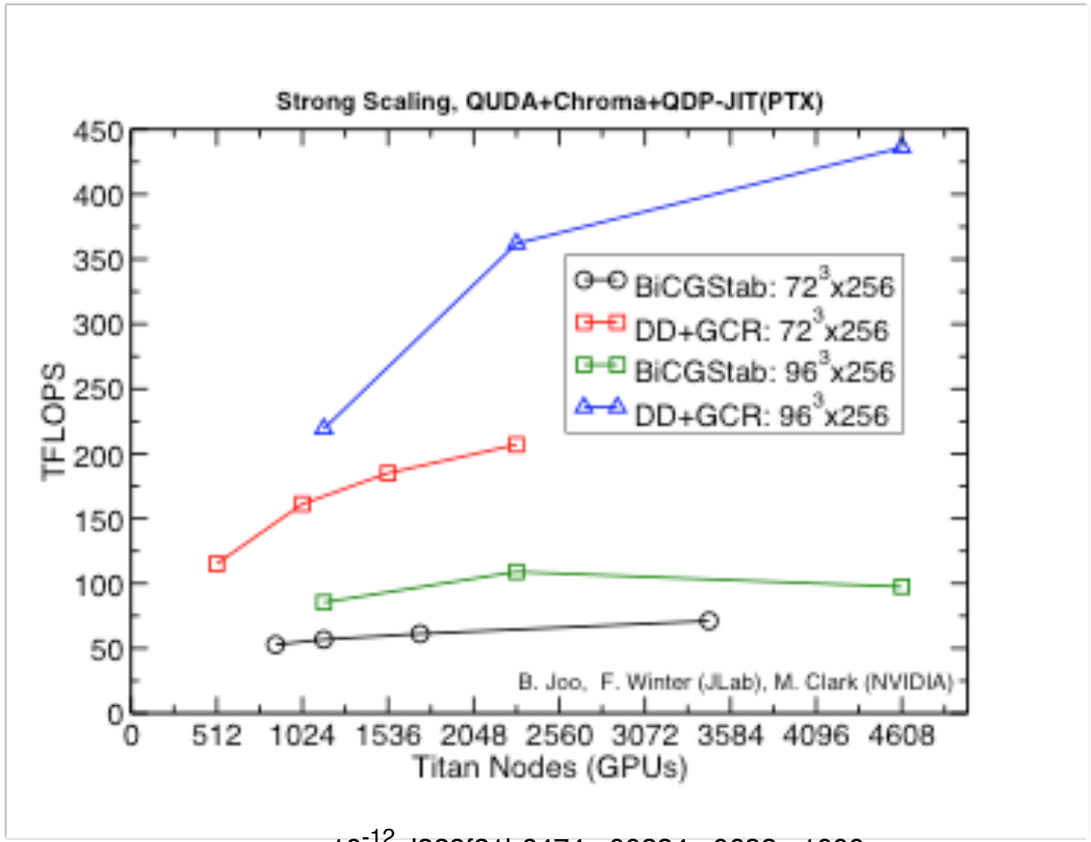
- Semi-implicit integrator: Repeated solution of Dirac equation + more for analysis.

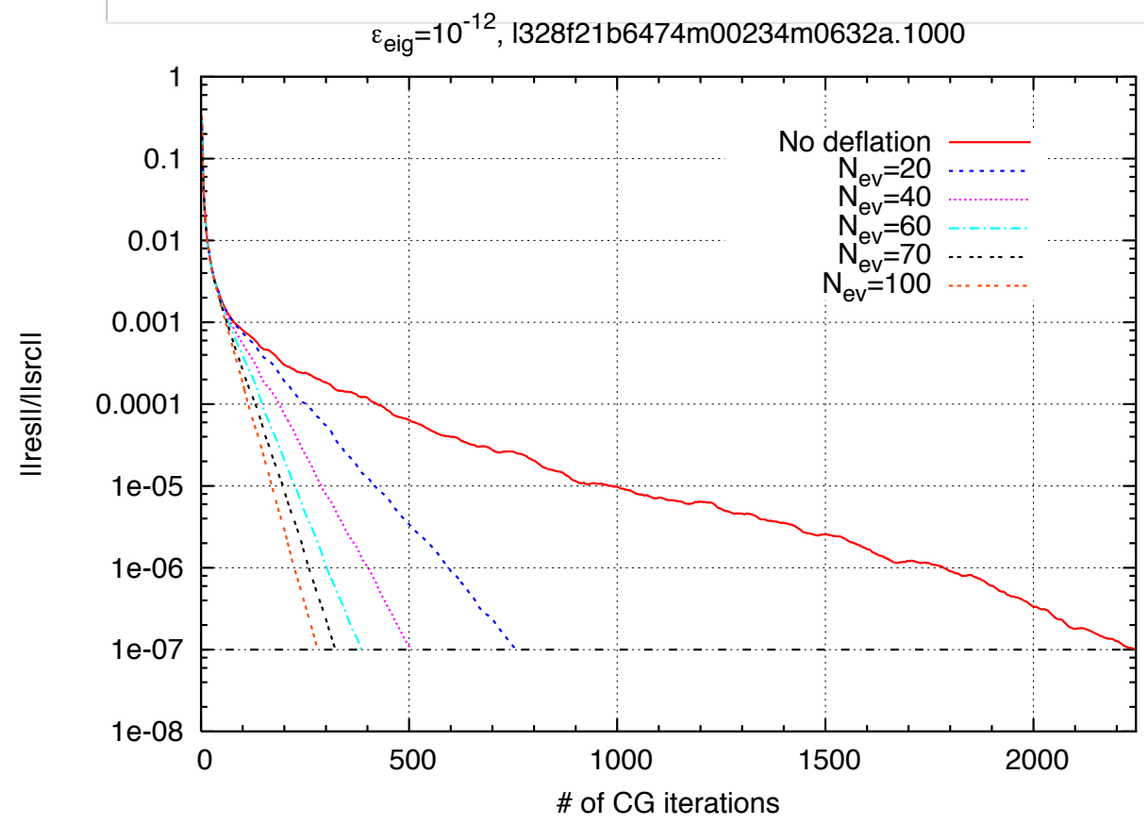


$$D_{quark} = m_q + \frac{1-\gamma_\mu}{2} U(x, x+\mu) + \frac{1+\gamma_\mu}{2} U(x+\mu, x)$$

Domain Decomposition & Deflation

- DD+GCR solver in QUDA
  - GCR solver with Additive Schwarz domain decomposed preconditioner
  - no communications in preconditioner
  - extensive use of 16-bit precision
- 2011: 256 GPUs on Edge cluster
- 2012: 768 GPUs on TitanDev
- 2013: On BlueWaters
  - ran on up to 2304 nodes (24 cabinets)
  - FLOPs scaling up to 1152 nodes
- Titan results: work in progress





Rough “state of the art” on space-time lattice Simulations

$$a(lattice) \ll 1/M_{proton} \ll 1/m_\pi \ll L(box)$$
$$0.06 \text{ fermi} \ll 0.2 \text{ fermi} \ll 1.4 \text{ fermi} \ll 6.0 \text{ fermi}$$

- Largest lattices now can reach physical pion mass!
- L^4 = 100x100x100x100 points (Ls = 16 Domain Wall)
- Discretize Gluon Field: 4xL^4 dense 3x3 complex matrices
- Discretize Quark Field: 24xL^4 by 24xL^4 sparse operator
- Use multi-time step semi-implicit Symplectic Hamiltonian Integrator:
- Each step many Dirac (Krylov) Linear Elliptic Solvers
- Average operators over large Ensemble of Gauge Field
- Extrapolate to get physics:
  - Continuum (a->0)
  - Space-time Volume (L->infty)
  - Chiral to small pion mass (m->small)

Lots of help from Applied Math and Physical Intuition

Many different people (TOPS, QCD) and institutions involved in the collaboration

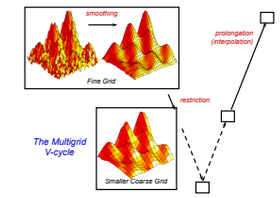
- CU Boulder
  - Tom Manteuffel
  - Steve McCormick
  - Marian Brezina
  - John Ruge
  - James Brannick
  - Christian Ketelsen
  - Scott MacLachlan
- Lawrence Livermore
  - Rob Falgout
- Columbia
  - David Keyes

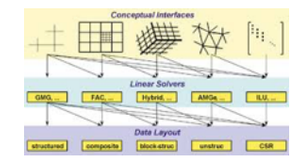
- Boston University
  - Rich Brower
  - Claudio Rebbi
  - Mike Clark
  - James Osborn
  - Saul Cohen
- Penn State
  - James Brannick
  - Ludmil Zikatanov
- Tufts
  - Scott MacLachlan
- Argonne
  - James Osborn

- NVIDIA
  - Mike Clark
  - Ron Babich
  - Michael Cheng
  - Oliver Witzel
- INT Seattle
  - Saul Cohen
- MIT
  - Andrew Pochinsky

FASTMath: Qlua+HYPRE

- QCD/Applied Math collaboration has long history: 8 QCDNA (Numerical Analysis) Workshops 1995-2014.
- Fast development framework is being constructed based on the combined strength of the FASTMath's HYPRE library at LLNL and the Qlua software at MIT.
- HYPER enhanced: Complex arithmetic and 4d and 5d hyper-cubic lattices.
- Qlua to HYPRE interface: to important Dirac Linear operators.
- Qlua is enhanced: 4d and 5d MG blocking and general "color" operators
- HYPRE: exploration of bootstrap algebraic multigrid (BAMG) algorithm for Dirac
- Goal to explore multi-scale for Wilson, Staggered and Dirac operators
- Test HYPRE methods at scale in Qlua and port into QUDA and QphiX libraries.





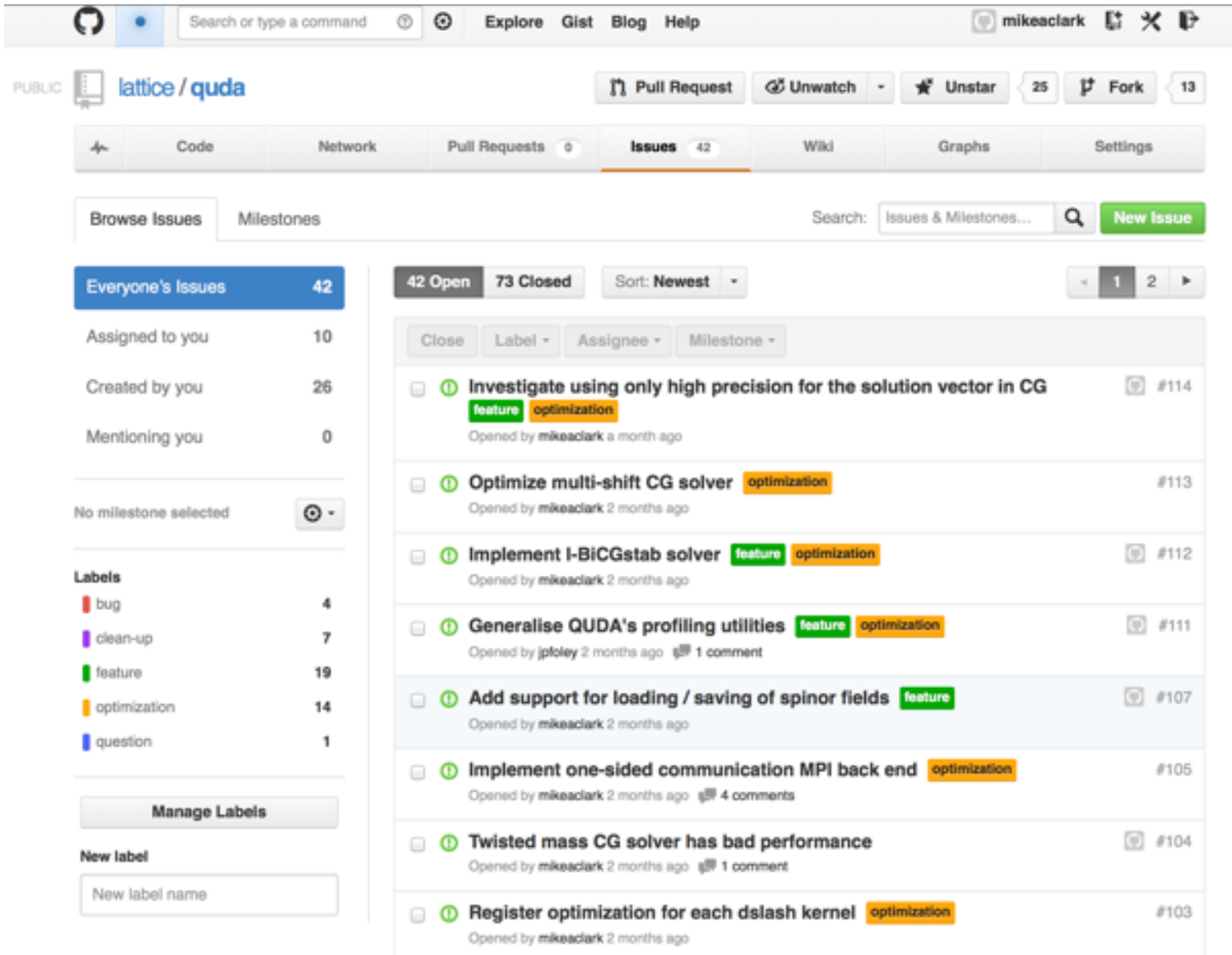
QUDA

QphiX

QUDA: NVIDIA GPU

- "QCD on CUDA" team – http://lattice.github.com/quda

- Ron Babich (BU-> NVIDIA)
- Kip Barros (BU ->LANL)
- Rich Brower (Boston University)
- Michael Cheng (Boston University)
- Mike Clark (BU-> NVIDIA)
- Justin Foley (University of Utah)
- Steve Gottlieb (Indiana University)
- Bálint Joó (Jlab)
- Claudio Rebbi (Boston University)
- Guochun Shi (NCSA -> Google)
- Alexei Strelchenko (Cyprus Inst.-> FNAL)
- Hyung-Jin Kim (BNL)
- Mathias Wagner (Bielefeld -> Indiana Univ)
- Frank Winter (UoE -> Jlab)



GPU code Development

- **REDUCE MEMORY TRAFFIC:**

  (1) Lossless Data Compression:

  SU(3) matrices are all unitary complex matrices with det = 1. 12-number parameterization: reconstruct full matrix on the fly in registers

$$\begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{pmatrix} \longrightarrow \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{pmatrix} \quad \mathbf{c} = (\mathbf{a} \times \mathbf{b})^*$$

- **Additional 384 (free) flops per site**

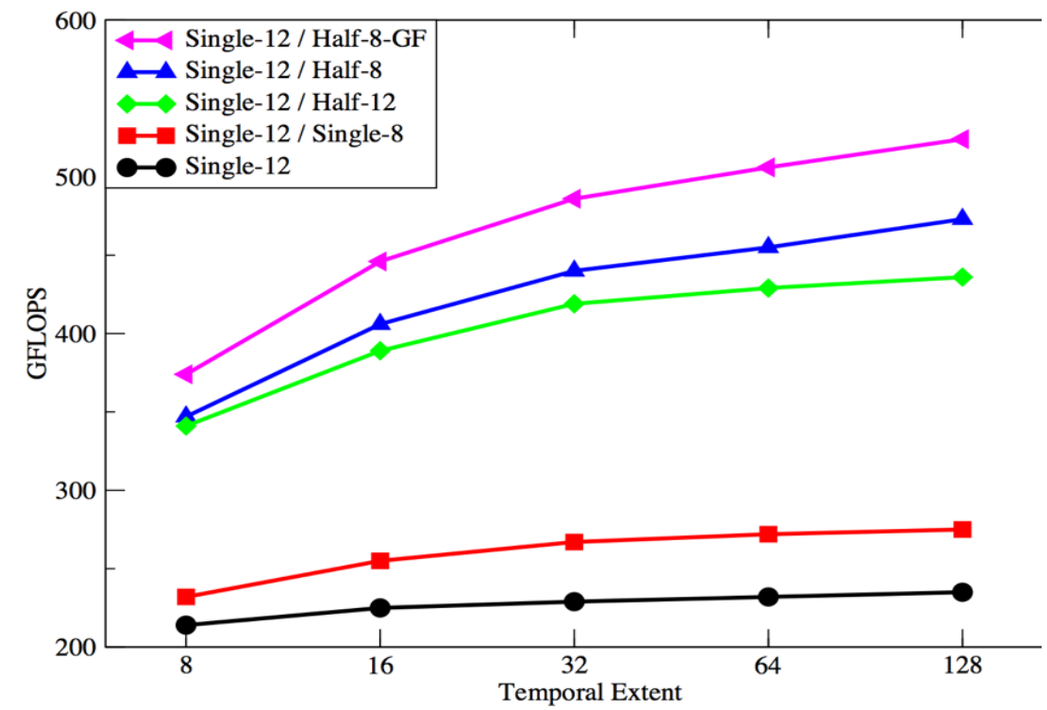
  Also have an 8-number parameterization of SU(3) manifold (requires sin/cos and sqrt)

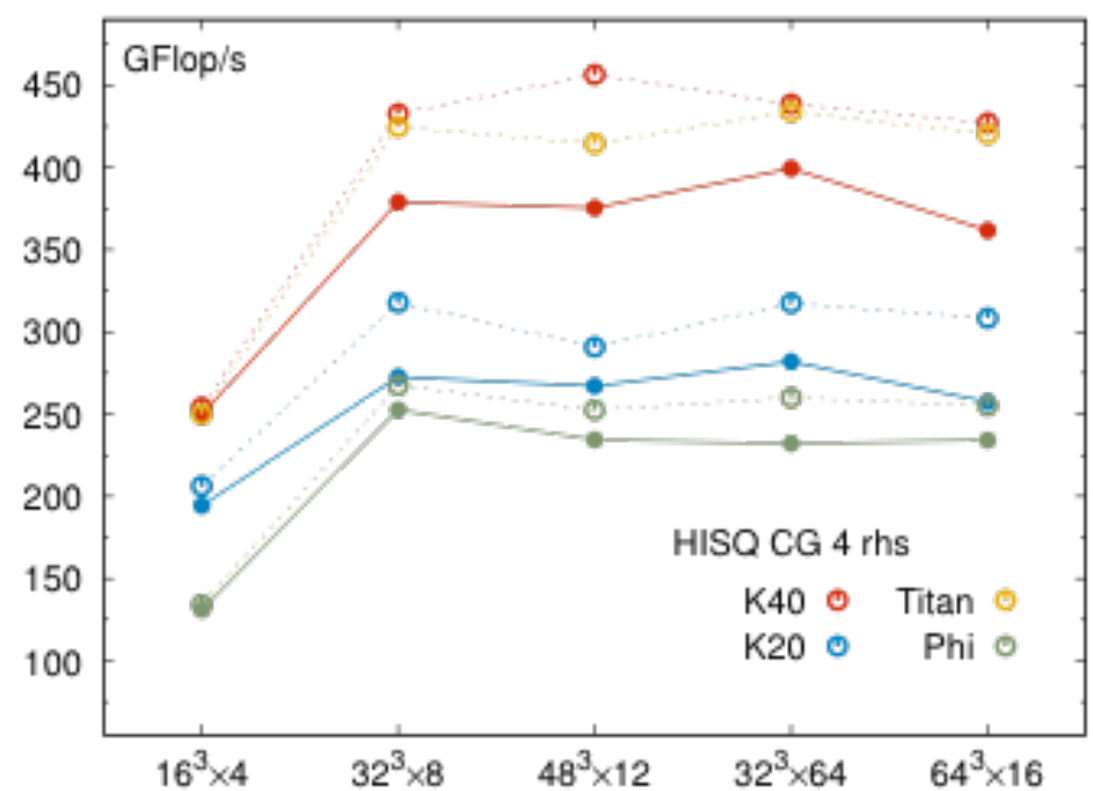
$$\text{Group Manifold}: S_3 \times S_5$$

  (2) Similarity Transforms to increase sparsity

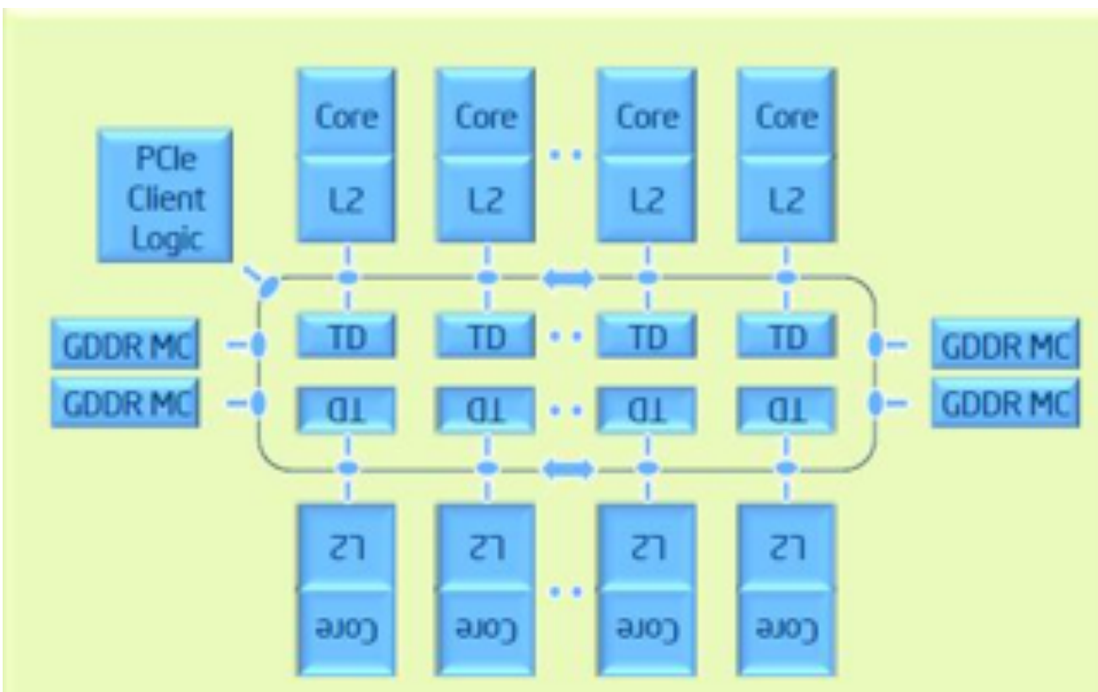
  (3) Mixed Precision: Use 16-bit fixed-point representation. No loss in precision with mixed-precision solves (Almost a free lunch:small increase in iteration count)

  (4) RHS: Multiples righthand sides





QphiX: Intel Xeon-Phi

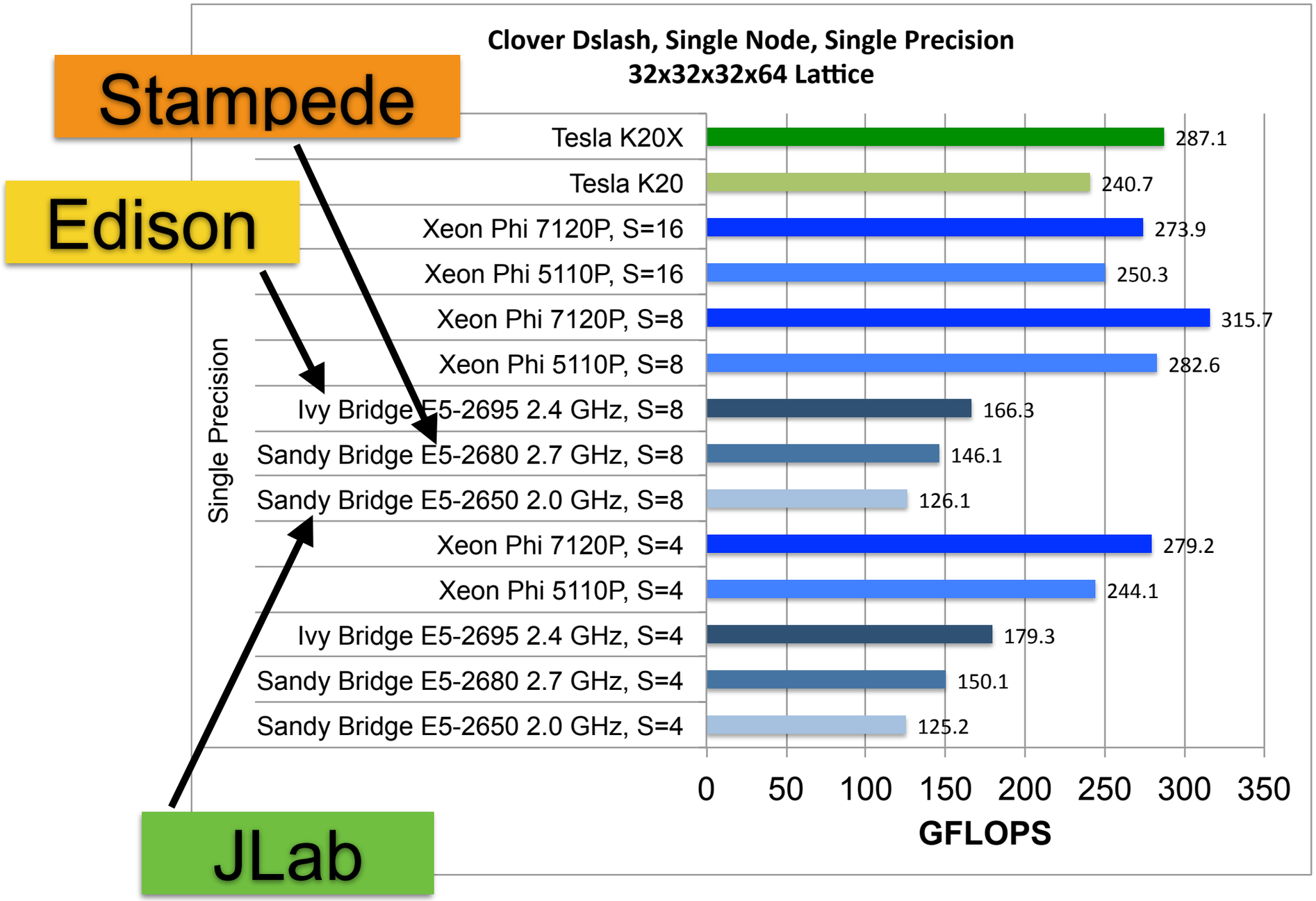


Xeon Phi 5110P

- 60 cores @ 1.053 GHz
  - connected by ring
  - 512Kb L2$ / core
  - 32KB L1I$ and 32KB L1D$
  - in-order cores, 4 way SIMT
  - 512 bit wide Vector Engine
    - 16 way SP/8 way DP
  - can do multiply-add
- Peak DP Flops: 1.0108 TF
- Peak SP Flops: 2.0216 TF
- 8 GB GDDR (ECC)
  - ‘top’ shows ~6GB free when idle

Source: http://software.intel.com/en-us/articles/intel-xeon-phi-coprocessor-codename-knights-corner
http://www.intel.com/content/www/us/en/processors/xeon/xeon-phi-detail.html

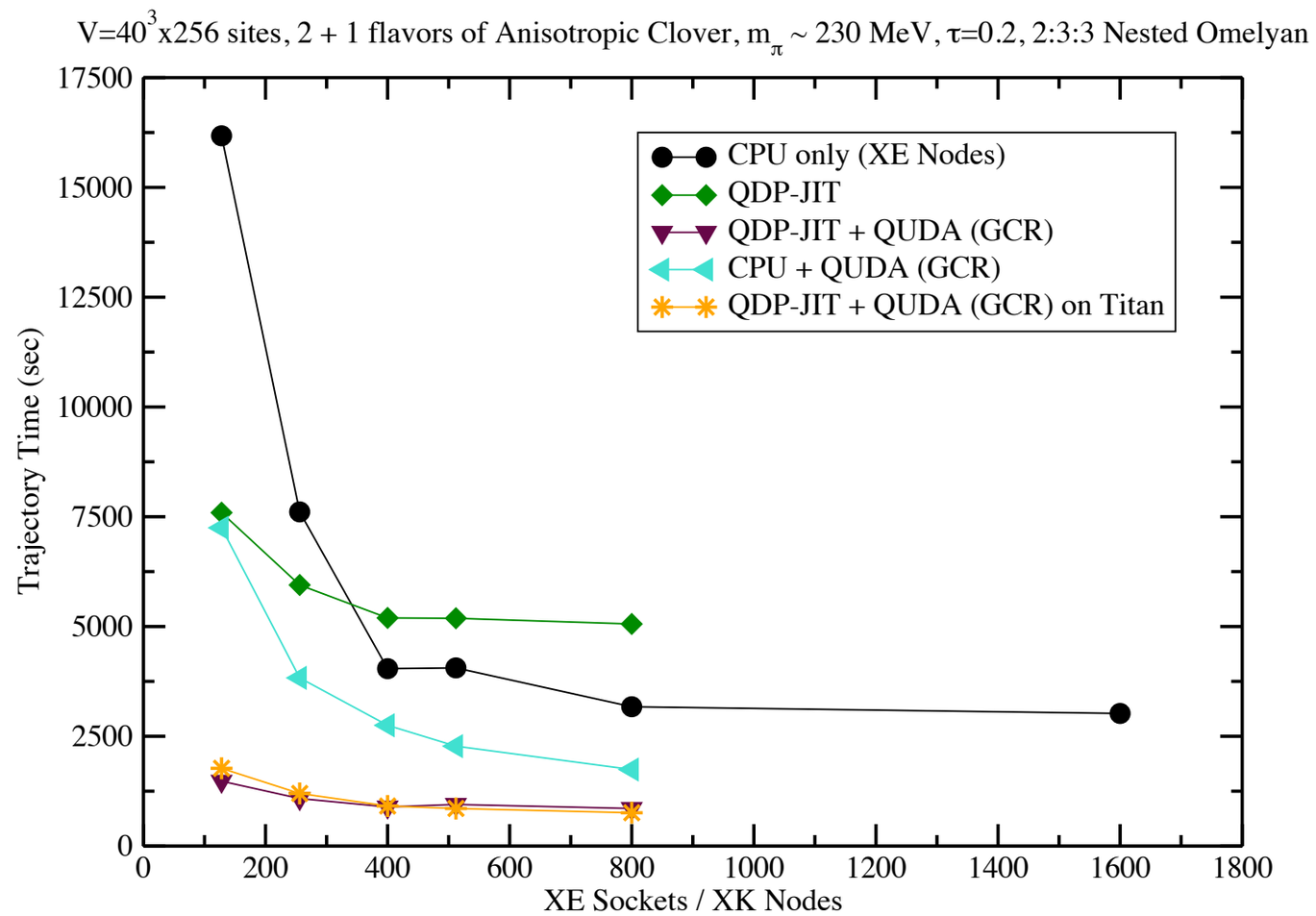
Xeon Phi and x86 Optimization



Performance of Clover-Dslash operator on a Xeon Phi Knight's Corner and other Xeon CPUs as well as NVIDIA Tesla GPUs in single precision using 2-row compression. Xeon Phi is competitive with GPUs. The performance gap between a dual socket Intel Xeon E5-2695 (Ivy Bridge) and the NVIDIA Tesla K20X in single precision is only a factor of 1.6x.

Gauge Generation using QDP-

- At 128 nodes QDP-JIT+QUDA outperforms old CPU code by about a factor of 11x
- At 800 nodes this decreases to about 3.7x
- Outperforms CPU+QUDA by between 5x - 2x (on 128 nodes and 800 nodes respectively)
- For this global volume (40^3x256) sites “shoulder region” is entered around 400 GPUs - limit strong scaling beyond that



F. T. Winter, M.A. Clark, R. G. Edwards, B. Joo, “A Framework for Lattice QCD Calculations on GPUs”, To appear at IPDPS’14, + Titan Data for QDP-JIT+QUDA(GCR)