Lattice Field Theory
Strong Dynamics in Standard Model and Beyond

SciDAC-3 PI Meeting

Richard Brower -- Boston University
(USQCD HEP& NP SciDAC software co-director)

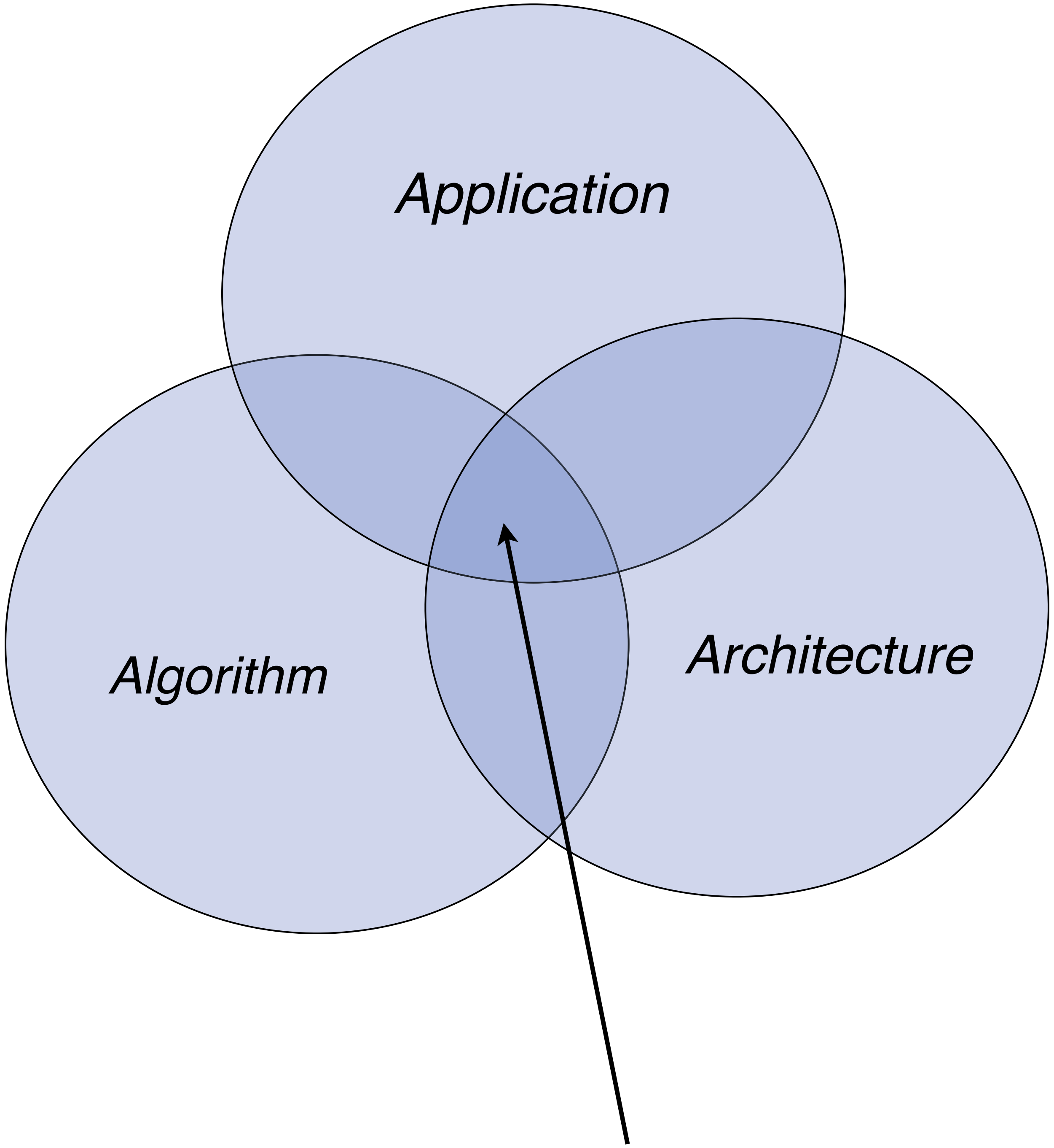
July 23, 2015

3 Part USQCD HEP Program

- PRECISION PHYSICS

- Multi-scale ALGORITHMS

- Parallel SOFTWARE/HARDWARE

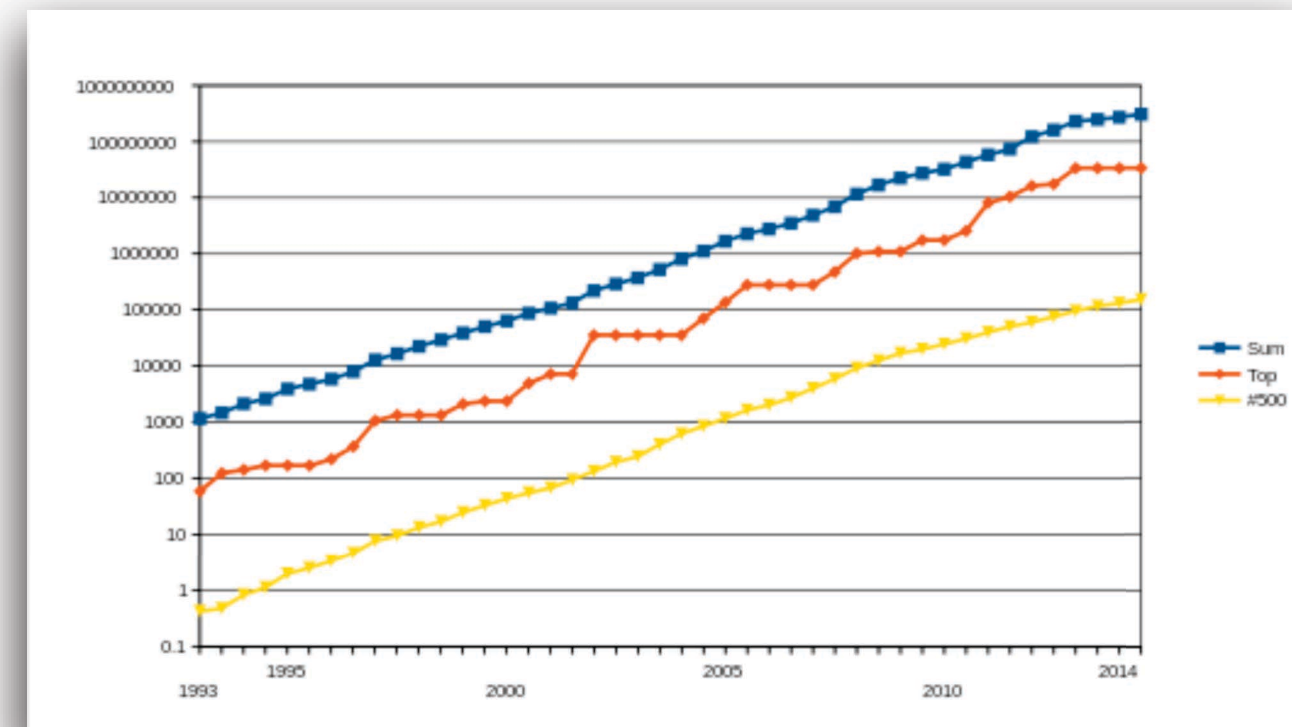


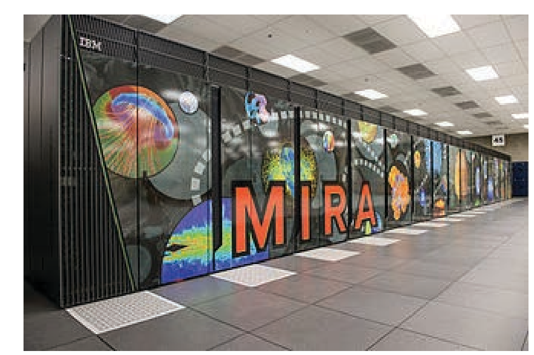
Need to find Sweet Spot

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?

Algorithms: Rapid (spectacular?) Evolution.

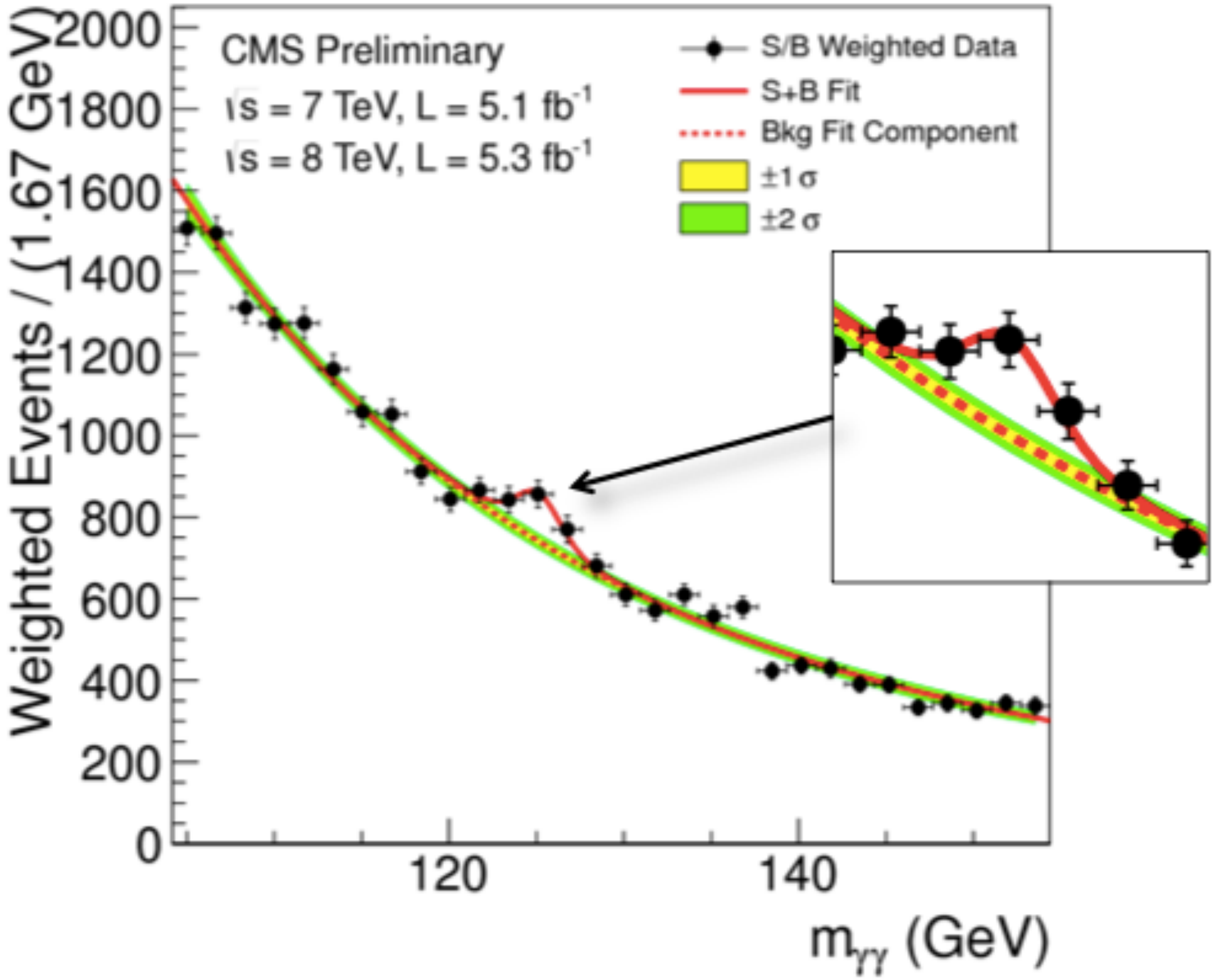
- Higher Resolution Physics exposes multi-scale
  - * Physical pion mass: u,d,s,c physical quark masses
  - **Multigrid for Domain Wall and Staggered solves on GPUs

- Heterogeneous Architecture requires data locality & communication reductions
  - Domain Decomposition for strong scaling
  - Multiple precession: half -> single -> double solution

- Huge opportunity and challenge but requires a long development software cycle:
  - --> algorithmic discovery (math)
  - --> full scale testing/optimization (computer science)
  - --> tuning to target architecture in production codes. (software eng/physics)

see HEP posters: * K-> 2pi by BNL/ANL & QUDA/Multigrid by FNAL/BU

Physics: Post Higgs Era

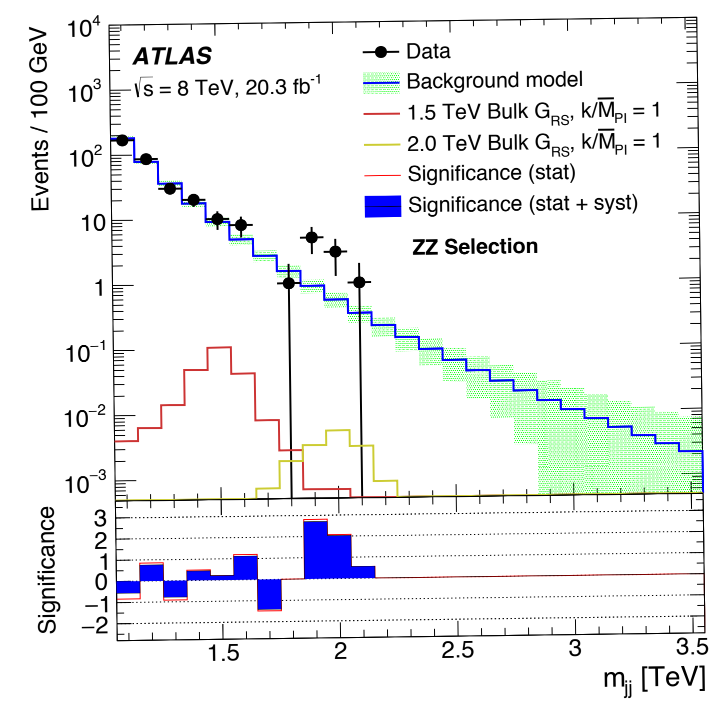


BUT still leaves us with the suspicions that some more fundamental physics lies beyond.

The SM with 26 free parameters has just too many “epicycles”.

We await the “Beyond the Standard Model” (BSM) Copernican Revolution.

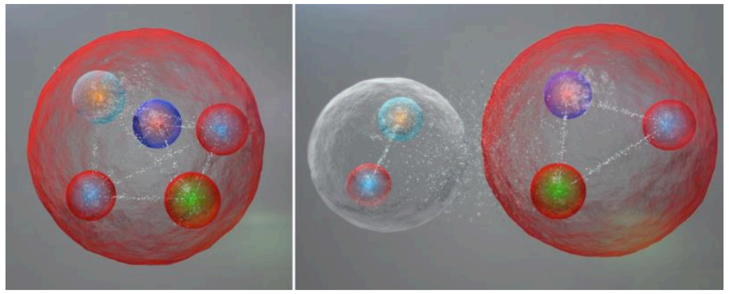
New BSM resonances?



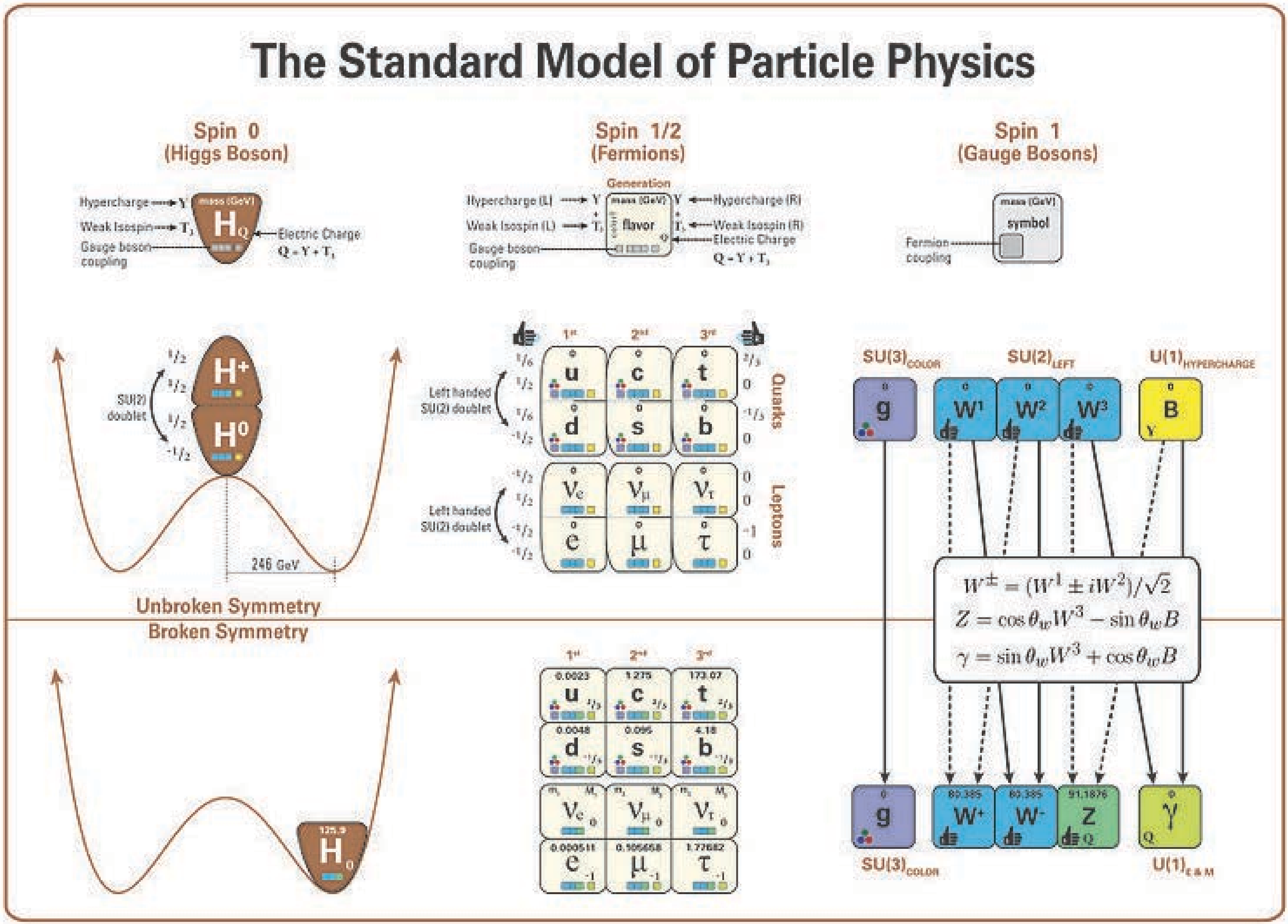
The difficulty of unravelling hints of BMS physics calls for “all hands on deck”:

Precision experiments must be matched step by step by increasing precision for QCD -- both high order Feynman expansions & high fidelity Lattice Field Theory.

Pentaquark states?



$26\ parameter\ Standard\ Model(SM)$

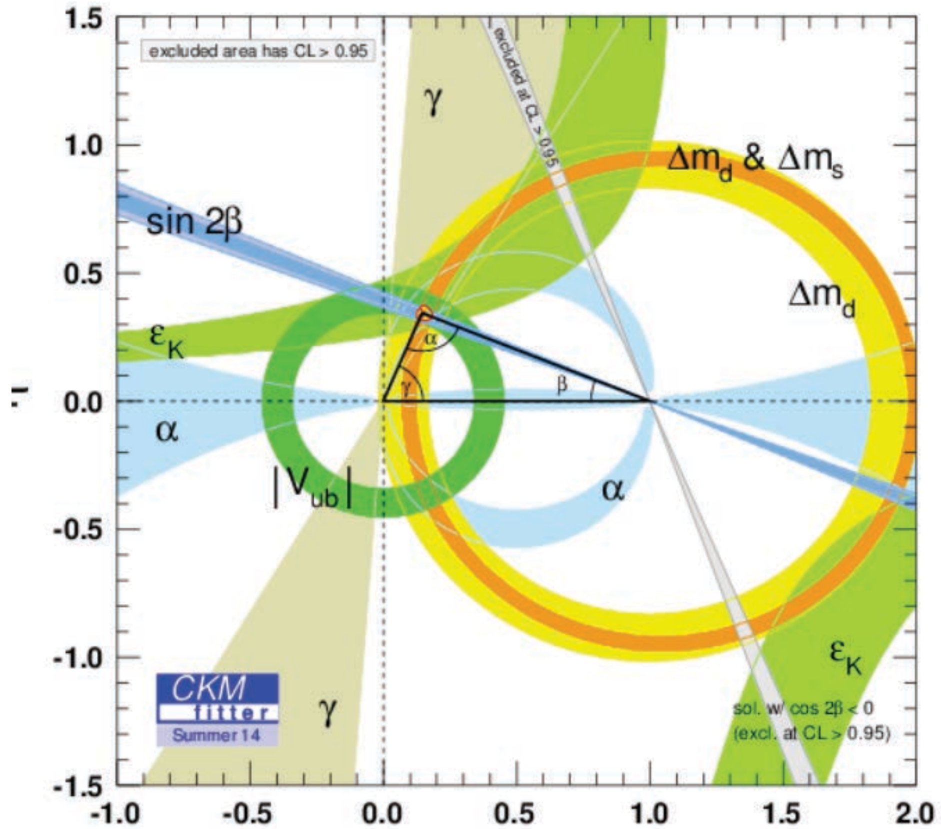


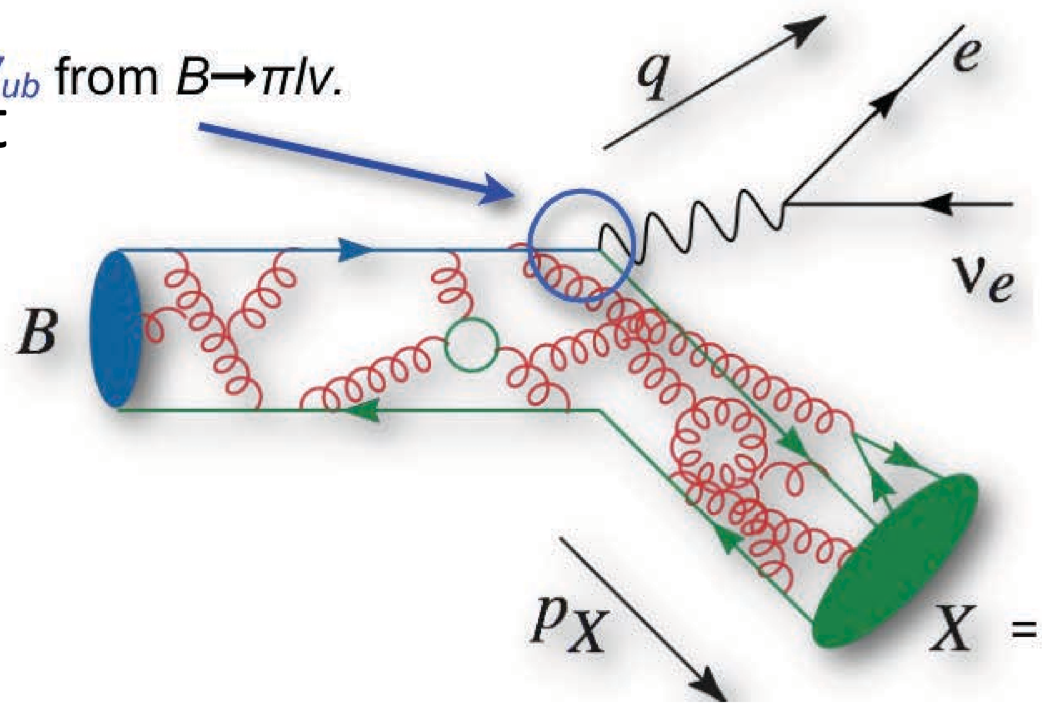
P5: Lattice role in HEP Physics Program

- P5 meeting found that "Lattice gauge theory will be needed ... almost everywhere"

- Nearly all physical process have higher order loop correction involving "quark loops" so QCD calculation are involved.

- See reported Paul Mackenzie at SciDAC PI 2014 for details





$B \to \pi l \nu$ semileptonic decay

High precision can be critical to discovery

- Classic examples
  - General Relativity: Perihelion of Mercury (1919)
  - Quantum Field Theory: Lamb Shift (1947)
  - No Klong --> 2 mu : Charm (Glashow, Iliopoulos, Maiani 1970)
  - Delta Mass of K --> Charm mass (Gaillard, Lee 1974)
- 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
  - mu->2e and mu->2gamma , neutrino scattering,
  - Dark matter detection through Higgs to Nucleus vertex
  - Composite Higgs and Dark Matter Theories?

QCD: Computational problem

What so difficult about this!

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

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

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

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

Lattice QCD formulation

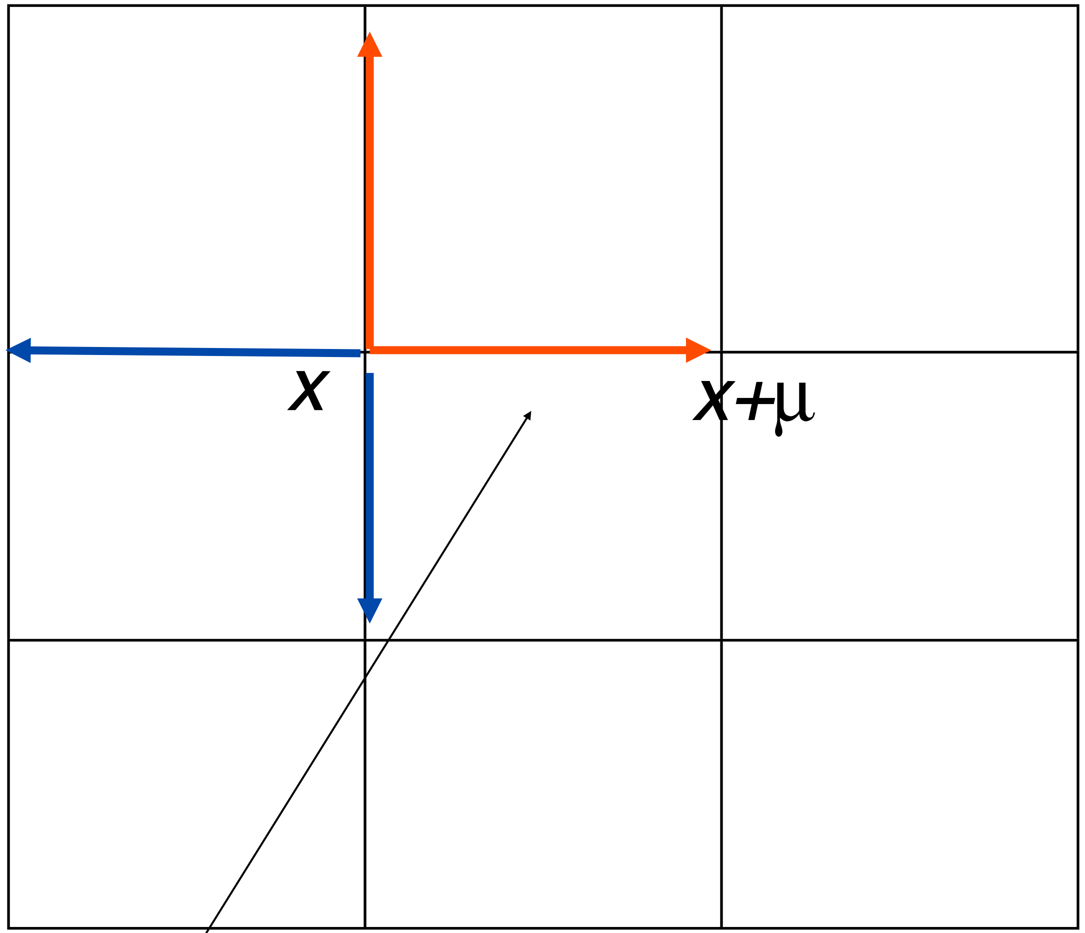
$$\mathrm{Prob}[U,\phi] = Z^{-1} e^{\beta Tr[U_{glue} + U^\dagger_{glue}] + \bar{\phi}(D^\dagger_{quark} D_{quark})^{-1}\phi}$$

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

- Get rid of Determinant with "pseudo-fermions"

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

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



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

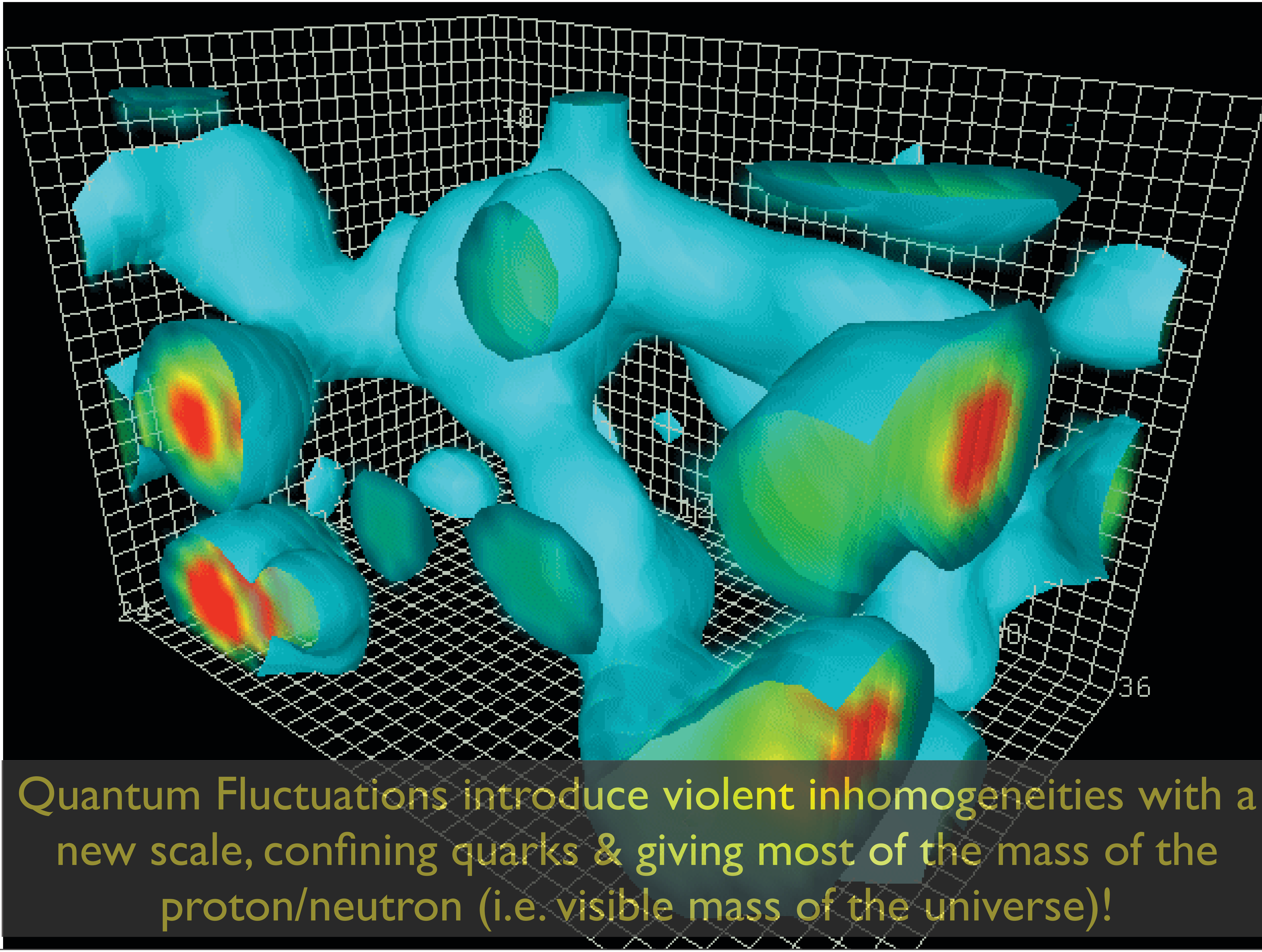
"A little knowledge is a dangerous thing"

Lattice Quantum Field Theory is NOT a typical applied 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	✓	✓	✗ **

QM spontaneously brakes symmetries causing (unexpected) large scales.

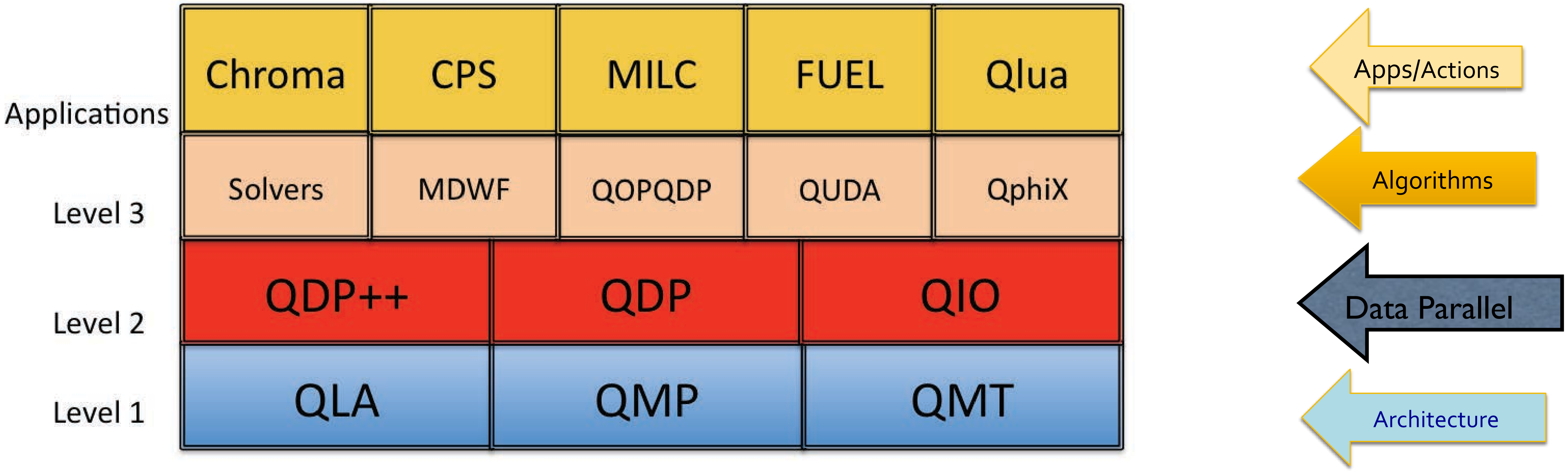
* color charge g is NOT a parameter! ** Small additional mass quarks (i.e. Higgs)



Quantum Fluctuations introduce violent inhomogeneities with a new scale, confining quarks & giving most of the mass of the proton/neutron (i.e. visible mass of the universe)!

USQCD Software Stack

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



Chroma = 4856 files

Wilson clover

CPS = 1749 files

Domain Wall

(full chiral sym)

MILC = 2300 files

Staggered

(partial chiral sym)

QUDA/python = 221 files

QLA/perl = 23000 files

Major USQCD Software contributors 2012-15

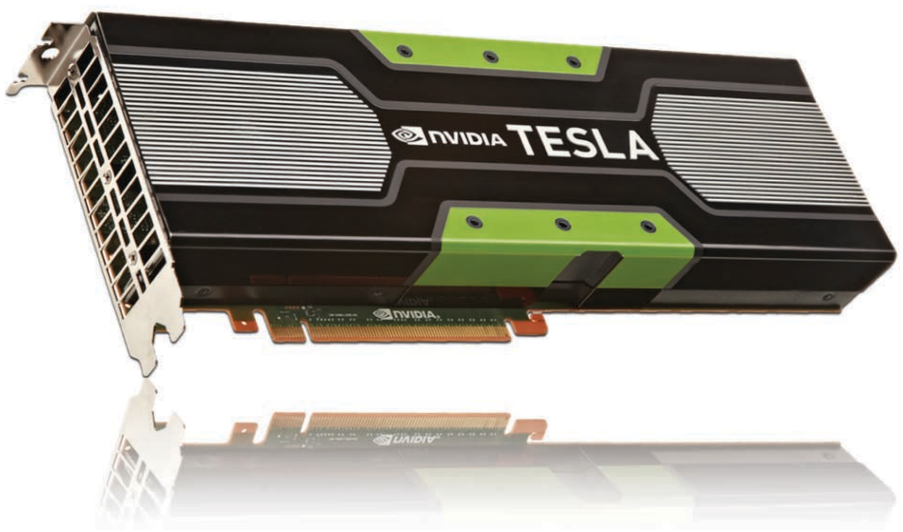
- ANL: James Osborn, Meifeng Lin (now at BNL) Heechang Na
- BNL: Frithjof Karsch, Chulwoo Jung, Hyung-Jin Kim,S. Syritsyn,Yu Maezawa
- Columbia: Robert Mawhinney, Hantao Yin
- FNAL: James Simone, Alexei Strelchenko, Don Holmgren, Paul Mackenzie
- JLab: Robert Edwards, Balint Joo, Jie Chen, Frank Winter, 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
- Small Fraction with direct SciDAC support.
- "Team of Rivals" (Many others in USQCD and Int'l Community volunteer to help

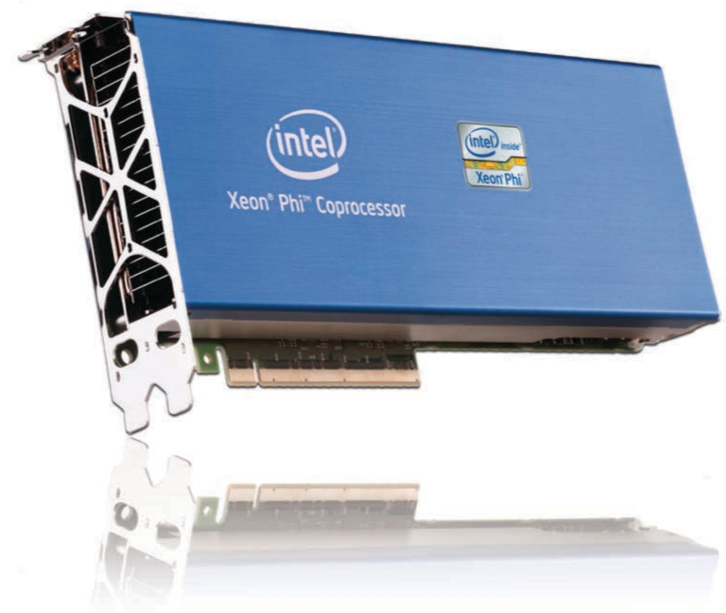
#1 PRIORITY: Physics Codes on INCITE & HPC.

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!)

#2 PRIORITY : Specialized Code Libraries





QUDA: (QCD in CUDA)

http://lattice.github.com/quda

QphiX (see NP talk)

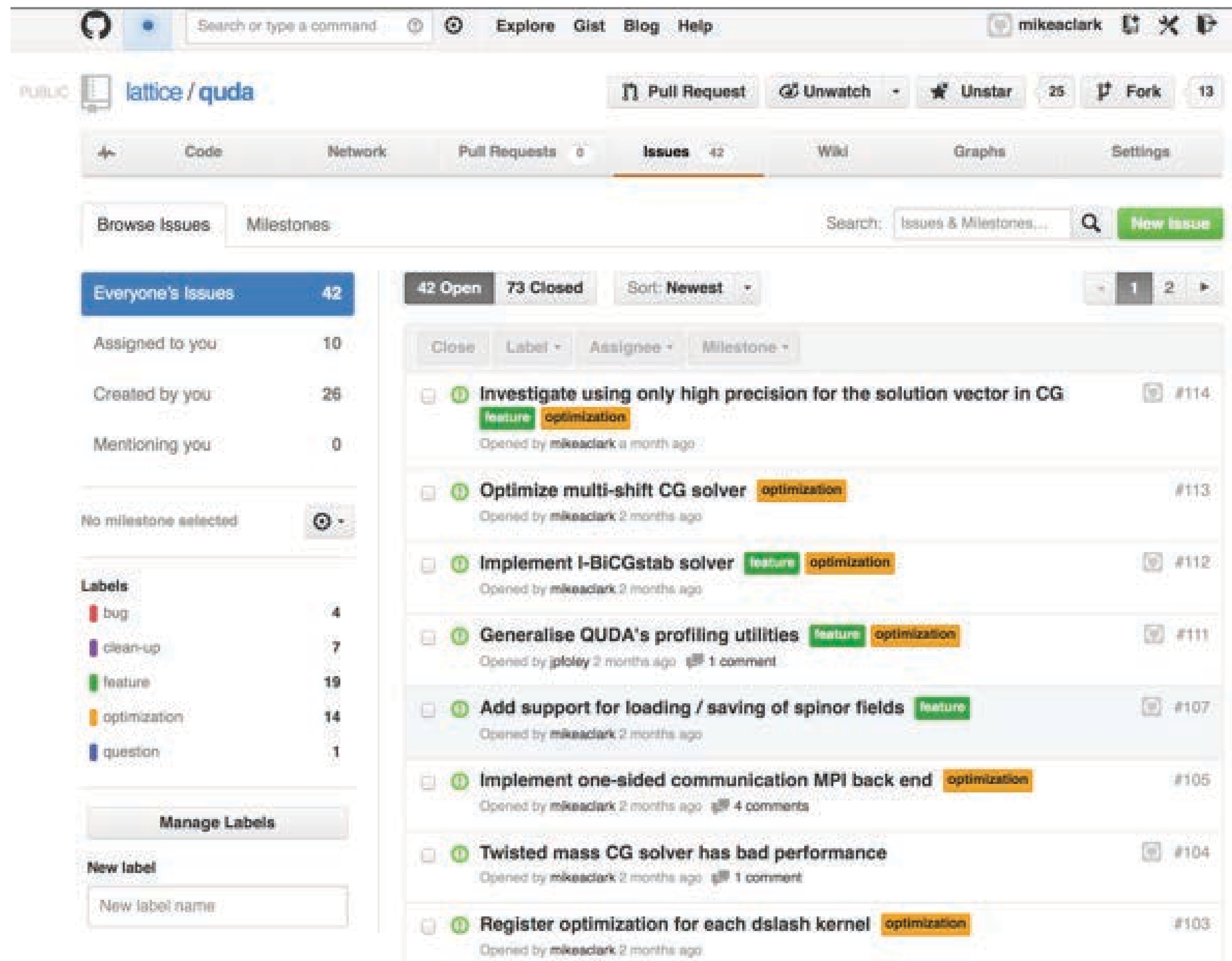
https://github.com/JeffersonLab/qphix

- Target CORAL at Oak Ridge (NVIDIA/IBM) and at ANL (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 IBM: Direct access to industry engineering and software professionals.

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**

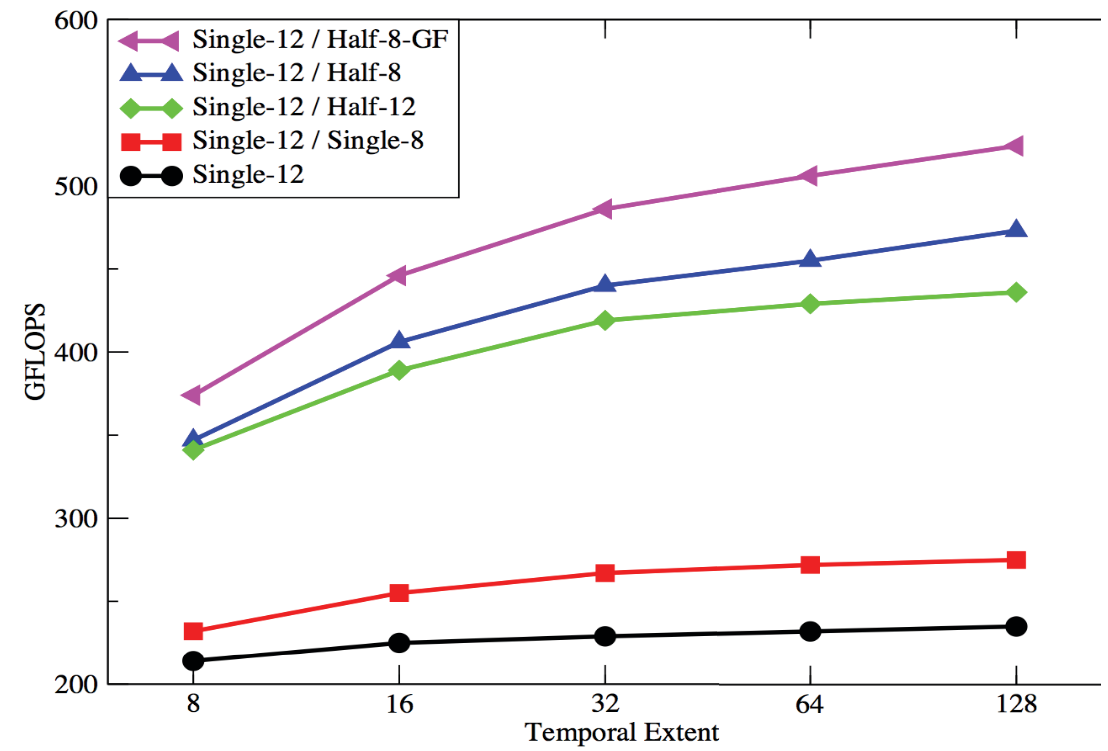
Also have an 8-number parameterization of SU(3) manifold (requires sin/cos and sqrt)

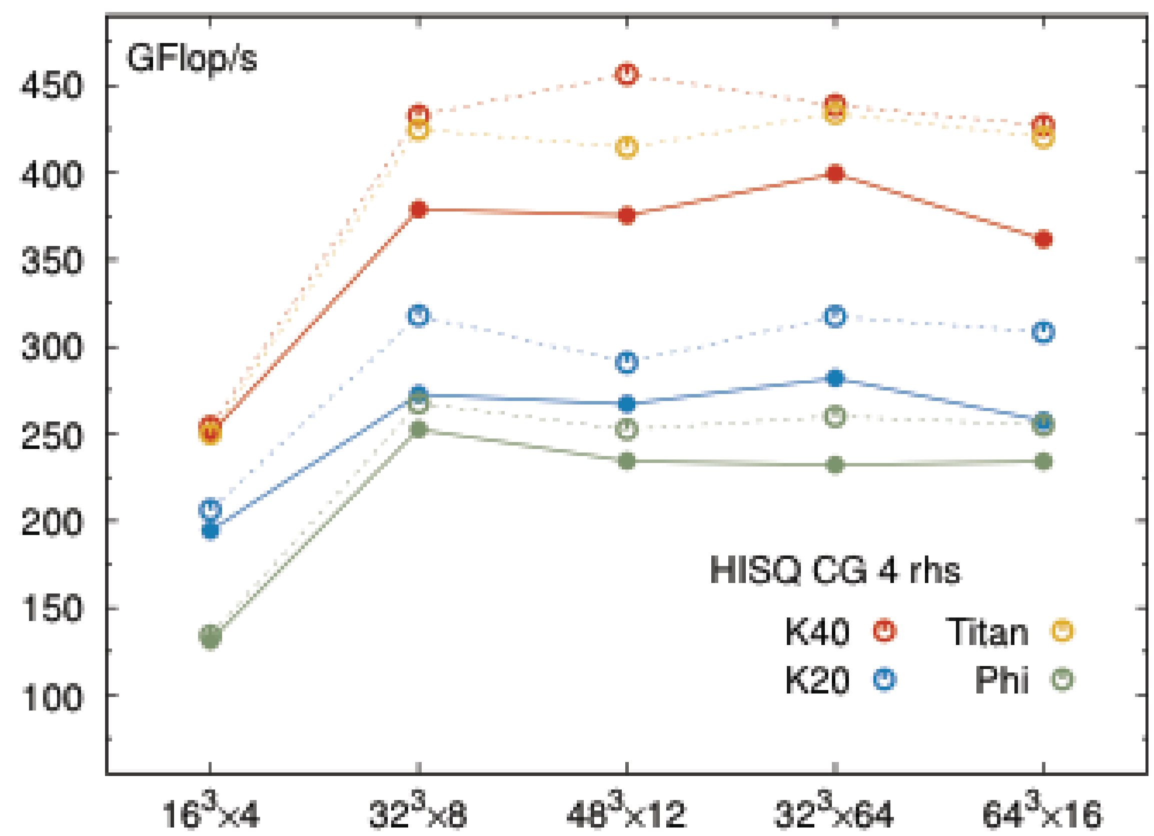
$$\text{Group Manifold:} S_3 \times S_5$$

(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





#3 PRIORITY: Algorithms (multi-scale et al)

- Multigrid Linear Solvers for 3 Lattice Dirac Actions.
  - Wilson Clover
  - Domain Wall or overlap (full chirality)*
  - Staggered (partial chirality or SUSY)*

- Hybrid Monte Carlo (HMC) Evolution
  - Multi-time step Symplectic Integration
  - Rational Forces Decomposition (RHMC)
  - Incorporate Multigrid for Rapid Equilibration and Evolution
  - zMoebius (Brower, Neff, Orginos + Blum, Izubuchi, Jung, Christoph and Syritsyn)

- Domain Decomposition for Communication Reduction
  - GCR solver with Additive Schwarz domain decomposed preconditioner

- Deflation Solvers & All Mode Averages for noise suppression

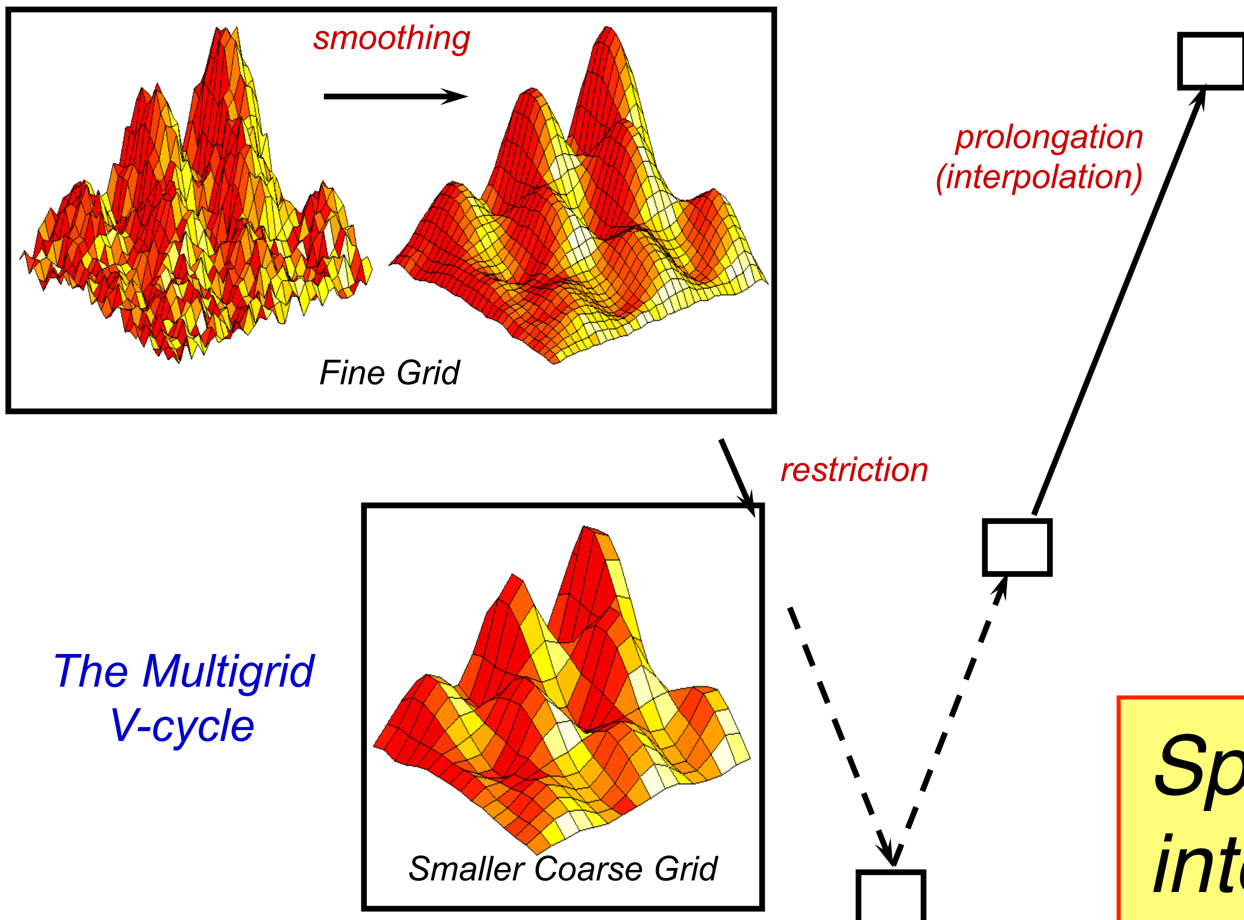
***Applications to Condensed Matter Lattice Theories! Nature has ubiquitous examples of Domain Wall and Staggered Fermions!**

‘Understanding Strongly Coupled Systems in High Energy and Condensed Matter Physics” Aspen May 24 -June 12, 2015

(Organizers: Brower, Catterall, Chandrasekharan, Sandvik, Scalettar, Wiese)

- Condensed Matter and HEP lattice field theory are both focussed on “strongly correlated fermionic systems”. There are promising common areas for collaboration both in theoretical and computational methods. (Report is being written)
  - Staggered and Domain Wall Fermions arise in many condensed matter systems. So fast solvers and study of role topologies and chirality represents a common concern.
  - New mass generation mechanism formulating lattice chiral gauge theories by mirror (Weyl) fermions in domain wall formulations or topological insulator models
  - Role of 4 fermi interaction in mass generation and composite Higgs models
  - The sign problem is if anything more prevalent for CM systems. Joint research here has potential. Relationship to entanglement.
  - Application Lagrangian path integral (or Hybrid Monte Carlo) methods and multigrid solvers to Graphene and similar lattice systems.

One Example: The Multigrid Solver (at last)



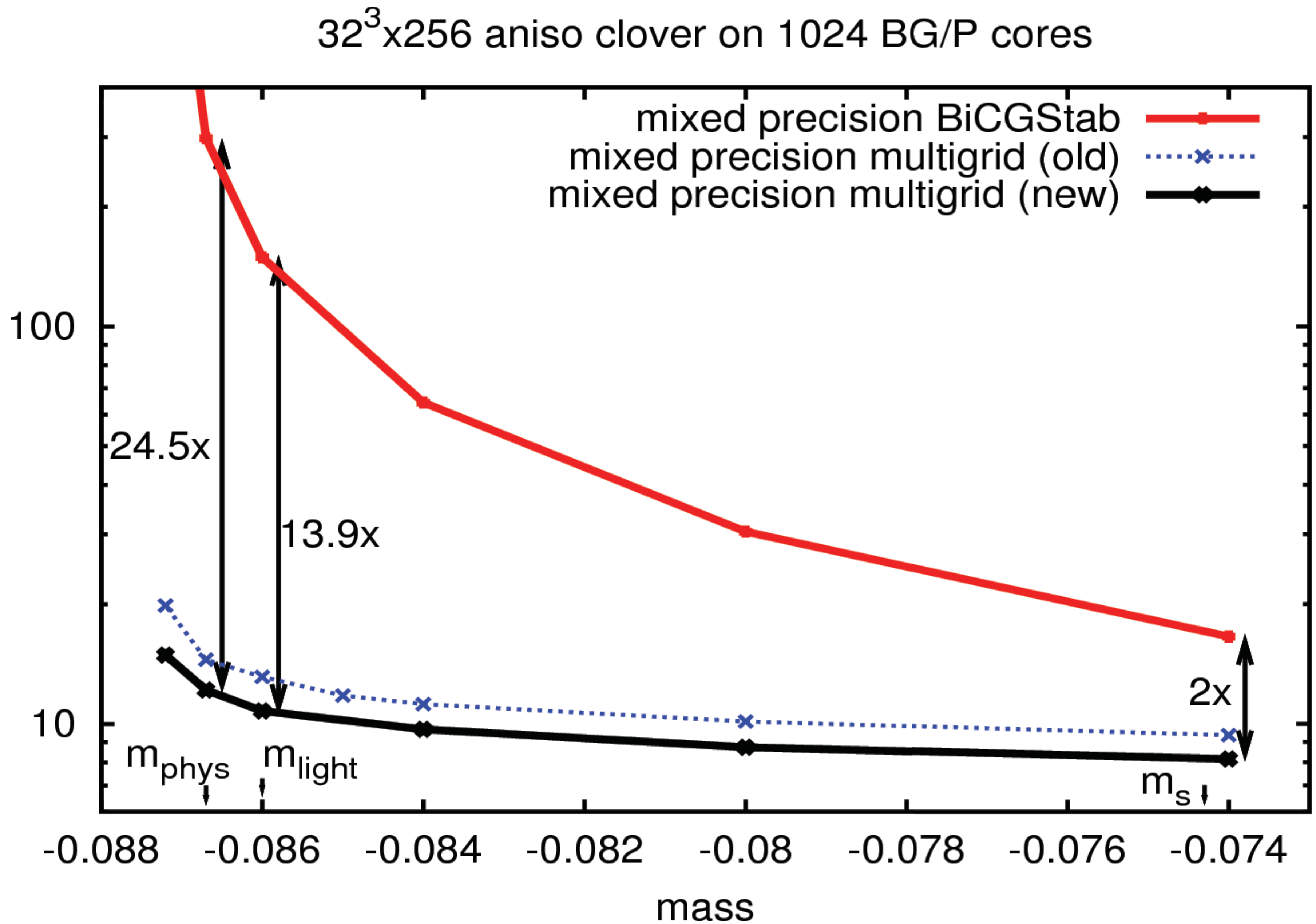
- 20 Years of QCD MULTIGRID
  - In 1991 Projective MG for algorithm to long distances (Brower, Edwards, Rebbi)
  - In 2011 Adaptive Geometric MG successfully extended

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

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

Slow modes are found by adapative "self learning" code: Near null space rich in low eigenvalue vectors.

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

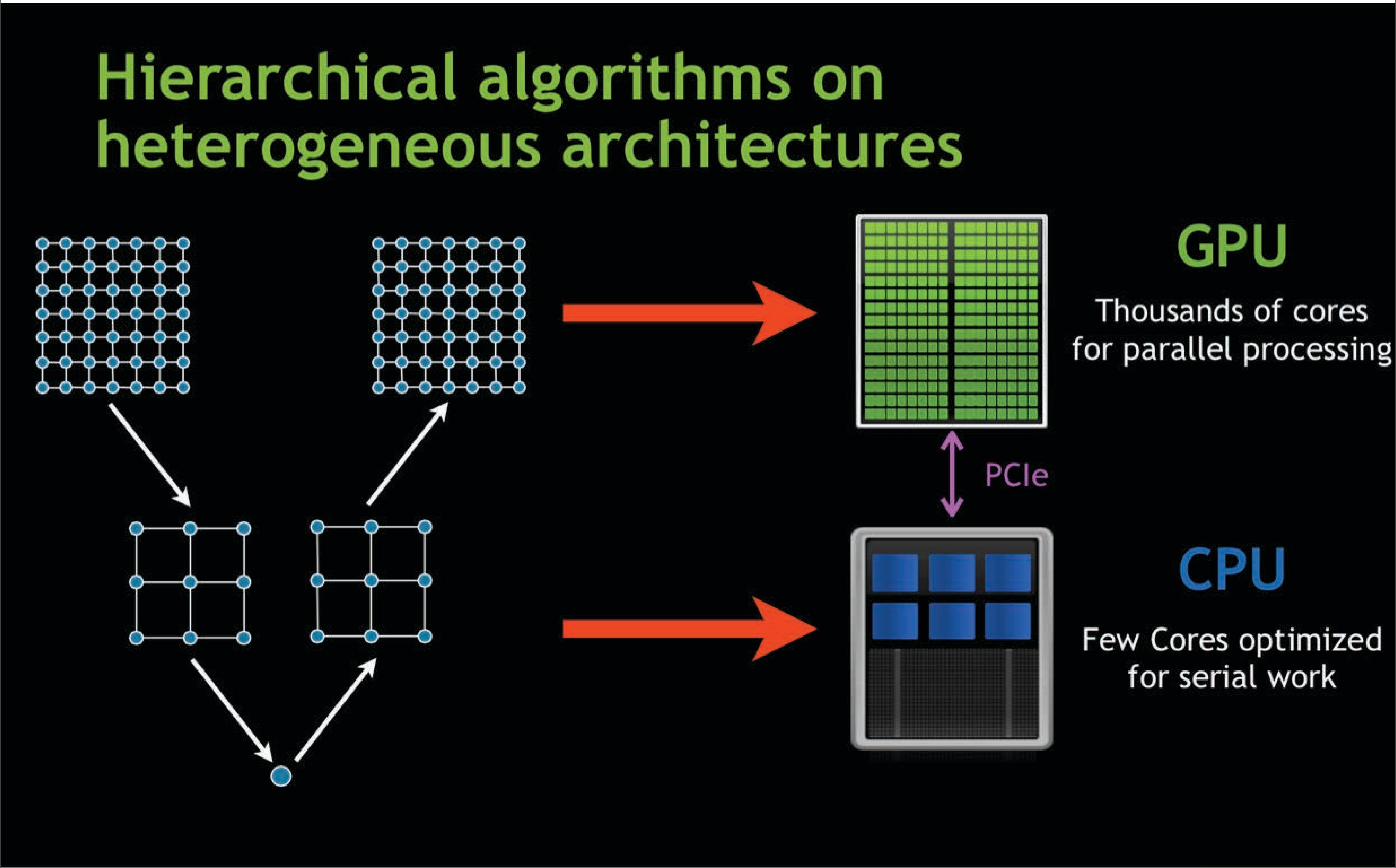


Adaptive Geometric Algebraic Multigrid

Performance on BG/Q at Argonne

"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

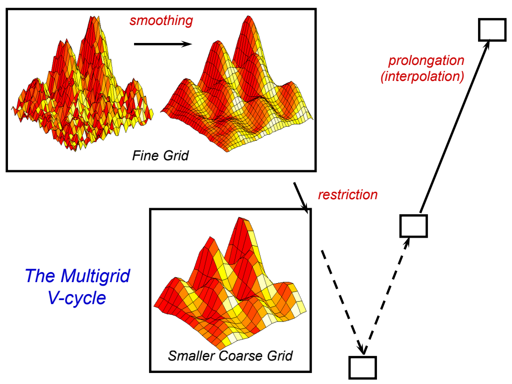


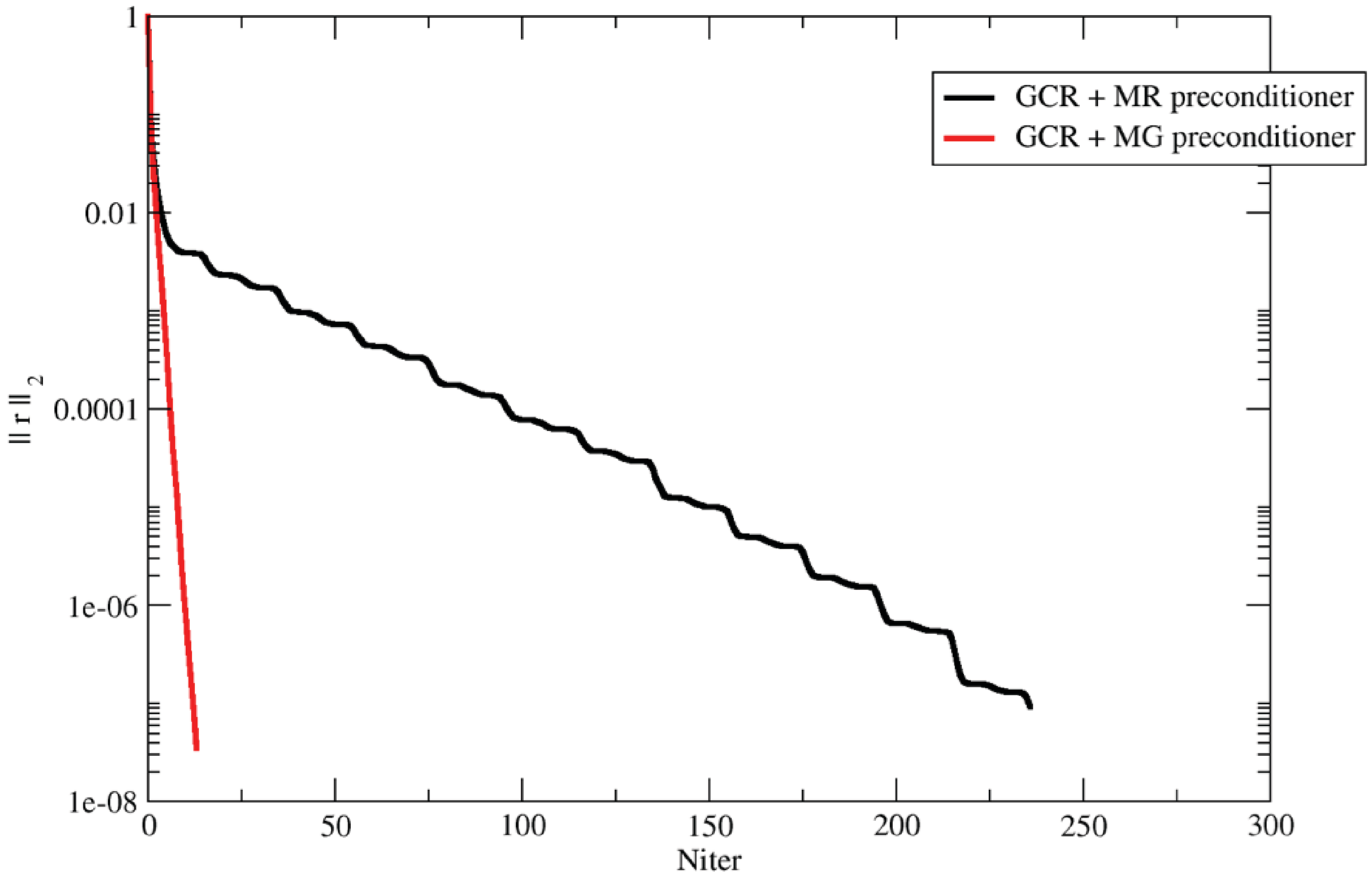
Multigrid on multi-GPU (then Phi):

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

Solution: Must put MG on GPU of course

+ =>





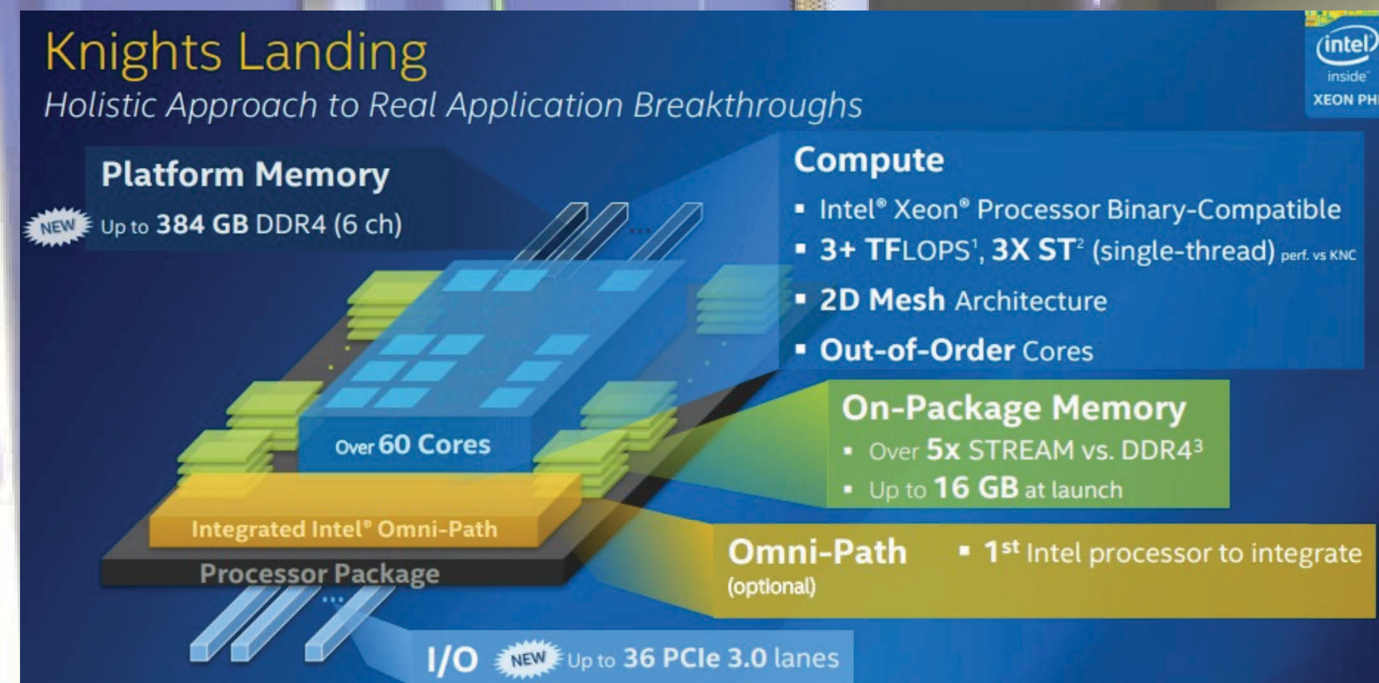
MG + GPU reduces cost $ by more than 100X

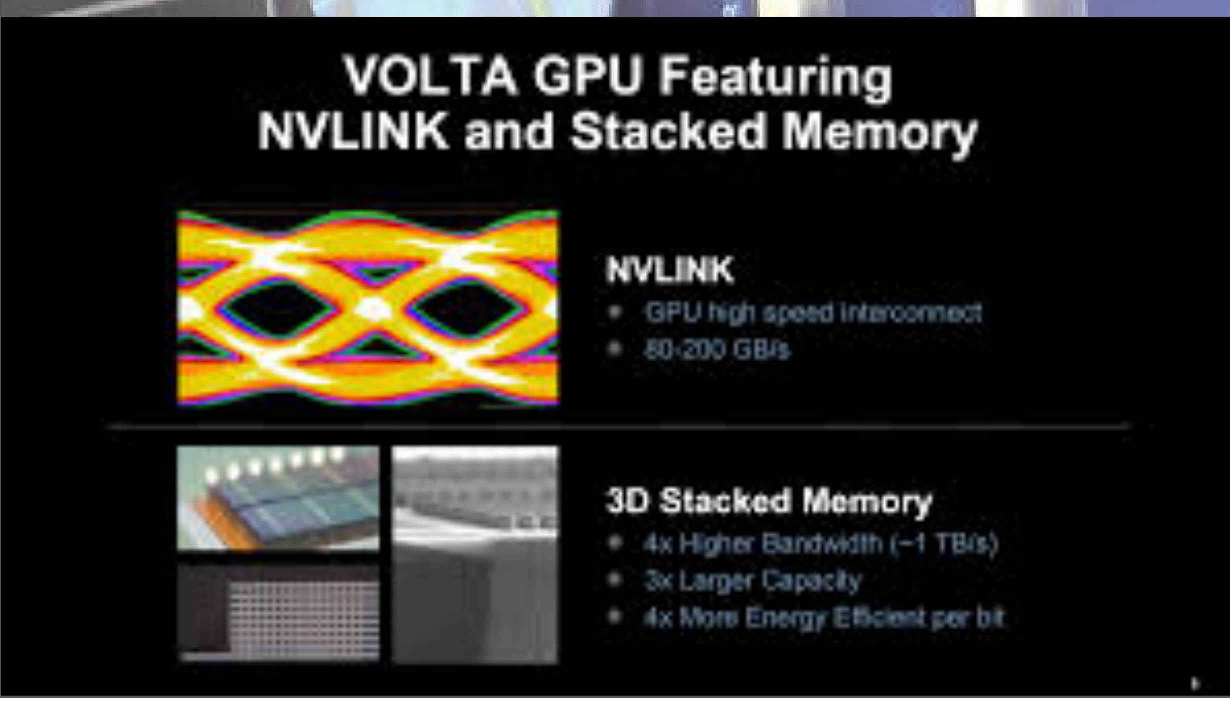
Near FUTURE: CORAL

Oak Ridge National Laboratory

NERSC

Argonne National Laboratory





	KNL Coprocessor	Host Processor	Host Processor with Integrated Fabric
Programming Model	Intel® 64 / AVX-512	Intel® 64 / AVX-512	Intel® 64 / AVX-512
I/O	PCIe	Fabric	Integrated Fabric
Power Efficiency	Baseline	>25% Better	>25% Better
Resiliency	Baseline	Intel server-class	Intel server-class
Performance	>3 TF	>3 TF	>3 TF
Memory Capacity	up to 16GB	up to 400GB	up to 400GB
Memory Bandwidth	>5x STREAM vs. DDR4	>5x STREAM vs. DDR4	>5x STREAM vs. DDR4

#4 PRIORITY: Framework for heterogeneous architecture

- Refactor our QCD API to keep pace with rapid evolution of heterogeneous architectures approaching the Exascale.
  - INTEL/CRAY (Knights Ferry/Corner/Landing/Hill..)
  - NVIDIA/IMB (Volta/OpenPower/NVLINK/Unified Memory)

- New Portable Data Parallel framework
  - FUEL (Lua Framework Argonne/James Osborn)
  - GRID (LBL/ Edinburgh/Peter Boyle/Chulwoo Jung)
  - Based on 3 parallels: MPI task/OpenMP thread/ SIMD vector

- We view these efforts for the next 2 years as an exploration of a new Data Parallel API. Comparison with performance of specialized libraries provides metric.

- Application to many lattice field theories: BSM beyond QCD, Conformal Field Theories (FEM), Condensed matter (e.g. Graphene) etc. Like CMSSL for the Thinking Machine of old!

$QUESTIONS$

Some Extra Slides

Designing multi-level algorithms for SciDAC QUDA library

HEP poster 1

R. Brower, M. Clark and A. Strelchenko

Boston University, Boston USA

NVIDIA

Fermi National Accelerator Laboratory, Batavia, Illinois, USA

1. Introduction

QUDA (QCD in CUDA) library started in 2008 with NVIDIAs CUDA implementation by Kip Barros and Mike Clark at Boston University. It has expanded to a broad base of USQCD SciDAC [1] software developers and is in wide use as the GPU backend for HEP and NP SciDAC application codes: Chroma, CPS, MILC, etc.

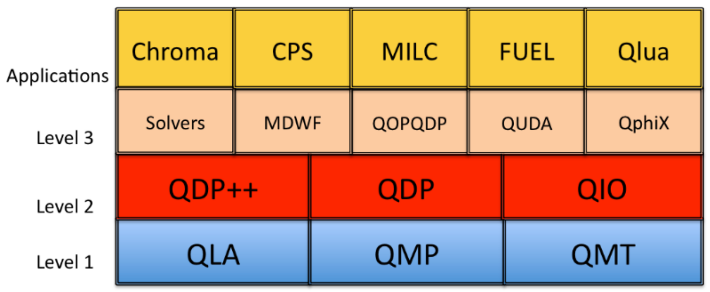
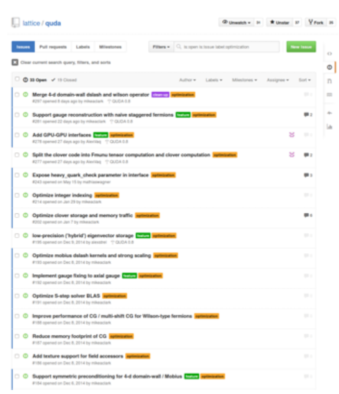


Figure 1: *SciDAC software stack.*

2. QUDA developers

- Ronald Babich (NVIDIA)
- Kipton Barros (LANL)
- Richard Brower (BU)
- Nuno Cardoso (NCSA)
- Michael Cheng (BU)
- Mike Clark (NVIDIA)
- Justin Foley (UU)
- Joel Giedt (RPI)
- Steven Gottlieb (IU)
- Balint Joo (JLab)
- Hyung-Jin Kim (Samsung)
- Claudio Rebbi (BU)
- Guochun Shi (NCSA)
- Alexei Strelchenko (FNAL)
- Alejandro Vaquero (INFN)
- Mathias Wagner (NVIDIA)
- Frank Winter (Jlab)



3. The QUDA library overview

QUDA is a library for performing calculations in lattice QCD on graphics processing units (GPUs), leveraging NVIDIA's CUDA platform. The current release includes optimized kernels for the following fermion operators:

- Wilson and Clover-improved Wilson
- Twisted mass (degenerate or non-degenerate, also with a clover term)
- Staggered and improved staggered (asqtad or HISQ) fermions
- Domain wall (4-d or 5-d preconditioned)
- Mobius fermions

Implementations of CG, multi-shift CG, BiCGstab, DD-preconditioned GCR and deflation algorithms (Lanczos and eigCG) are provided, including robust mixed-precision variants supporting combinations of double, single, and half (16-bit "block floating point") precision. The library also includes auxiliary routines necessary for Hybrid Monte Carlo, such as HISQ link fattening, force terms and clover-field construction. Use of many GPUs in parallel is supported throughout, with communication handled by QMP or MPI.

4. Hierarchical algorithms on heterogeneous architectures

- The multi-level framework:

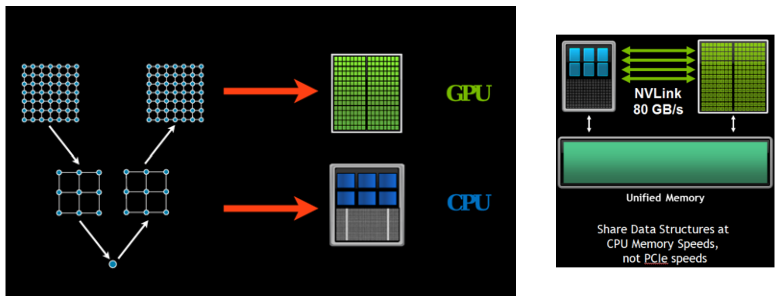


Figure 2: *Lattice multi-grid framework.*

- Adaptive Geometric Multigrid

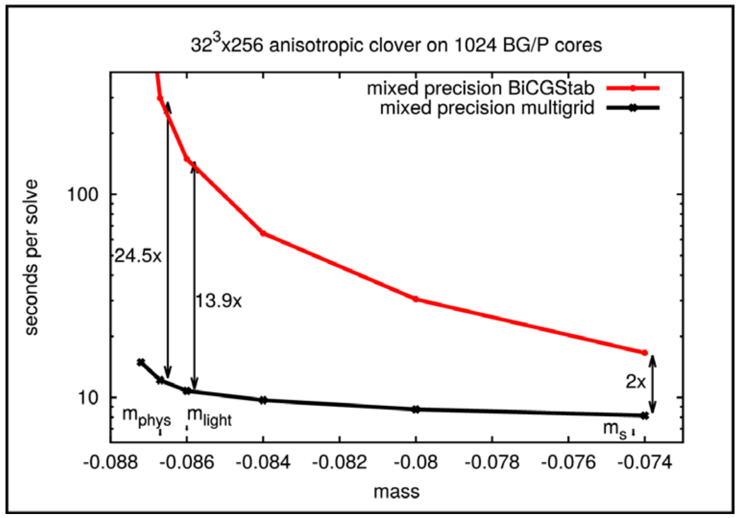


Figure 3: *Improving performance with the multigrid preconditioning.*

- Adaptive Geometric Multigrid on GPUs

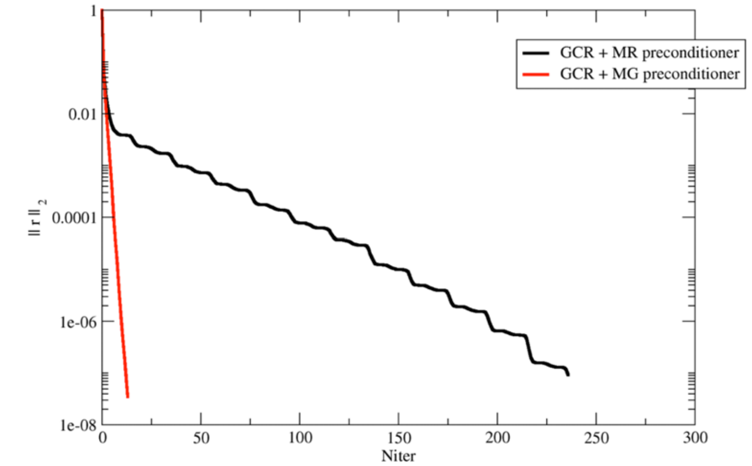


Figure 4: *Wilson multigrid on NVIDIA GPUs.*

5. Towards exascale computing

Figure 5: *New HPC facilities.*

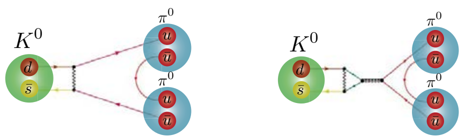
References

[1] M. A. Clark *et al*, Comput. Phys. Commun. **181**, 1517 (2010) [arXiv:0911.3191 [hep-lat]].

[2] R. Babich, J. Brannick, R. C. Brower, M. A. Clark, T. A. Manteuffel, S. F. McCormick, J. C. Osborn and C. Rebbi, Phys. Rev. Lett. **105**, 201602 (2010) [arXiv:1005.3043 [hep-lat]].

[3] http://lattice.github.io/quda/

Computational Advances and Precision Kaon Physics



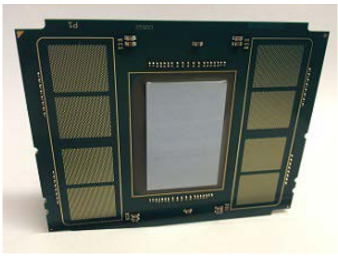
- Kaon decays to two pions are well measured experimentally.
- The Standard Model predicts the details of these decays, but requires knowledge of quark properties inside kaons/pions.
- Knowledge of quark properties is determined from first principles through Lattice QCD.

Lattice QCD simulations are now regularly being done with physical values for the light, strange and charm quarks. This allows many complicated phenomena to be investigated with direct impact on experiments. One example is the precision kaon physics being pursued by the RBC and UKQCD Collaborations, where the first calculation of direct CP violation in kaon decays has been done. In order to reduce the errors on this result, advances in algorithms and software are needed. Here we discuss the zMobius formulation of Domain Wall Fermions and the improvments it has led to to date. The CPS code is currently being evolved to utilize the Grid software of Peter Boyle, to be ready to exploit Intel's Knights Landing architecture when it becomes available early next year. An early target for our research is the reduction of the errors on this direct CP violation calculation as soon as the Knight's Landing machines become available.

SciDAC-3 HEP
Paul MacKenzie, PI

Robert Mawhinney
Columbia University

SciDAC-3 Principal Investigators Meeting
July 22-24, 2015



Knights Landing

HEP poster 2

Direct CP violation in kaons

- For kaons, the size of the violation of charge conjugation (particle ↔ antiparticle) and parity (physical process ↔ mirror image of process) symmetry (CP symmetry violation) is determined by the values for

$$\eta_{00} \equiv \frac{A(K_L \to \pi^0\,\pi^0)}{A(K_s \to \pi^0\,\pi^0)} = \epsilon - 2\epsilon'$$

$$\eta_{+-} \equiv \frac{A(K_L \to \pi^+\,\pi^-)}{A(K_s \to \pi^+\,\pi^-)} = \epsilon + \epsilon'$$

- Non-zero values are found by experiments, so CP violation exists:

$$|\epsilon| = (2.228 \pm 0.011) \times 10^{-3}$$

$$\mathrm{Re}(\epsilon'/\epsilon) = (1.65 \pm 0.26) \times 10^{-3}$$

- To relate ε to the Standard Model, quark properties in mesons are needed. These are contained in a parameter, called B_K, while all other terms in this formula can be measured independently.

$$\varepsilon_K = \kappa_\varepsilon\, C_\varepsilon\, \hat{B}_K\, \mathrm{Im}(\lambda_t)\, \{\mathrm{Re}(\lambda_c)[\eta_1\, S_0(x_c) - \eta_3\, S_0(x_c, x_t)] - \mathrm{Re}(\lambda_t)\, \eta_2\, S_0(x_t)\}\, e^{i\pi/4}$$

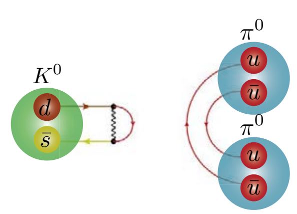
- B_K is currently known to 1.3% accuracy, after many years of lattice QCD work (Eur. Phys. J. C74 (2014) 2890).

$$\hat{B}_K = 0.7661(99)$$

- Many other lattice calculations, particularly involving heavy quarks, are also used to relate different Standard Model parameters to experiment.
- A first complete calculation for ε' has recently been completed by the RBC and UKQCD Collaborations (arXiv:1505.07863) yielding

$$\mathrm{Re}(\epsilon'/\epsilon) = (0.138 \pm 0.515_{\mathrm{stat}} \pm 0.443_{\mathrm{sys}}) \times 10^{-3}$$

- This result is slightly over 2σ from the experimental value
- Many important theoretical physics techniques are required in this calculation and the value is sensitive to physics beyond the Standard Model
- The biggest numerical hurdle is the calculation of disconnected quark diagrams, which are only connected only by gluons and are statistically noisy

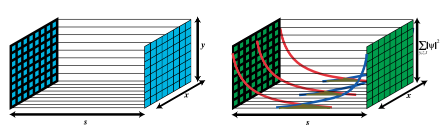


Algorithmic and Software Improvements

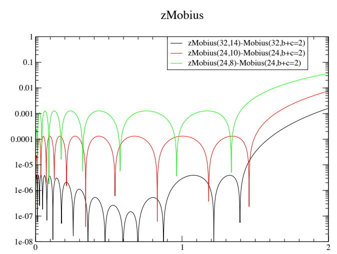
- These direct calculations of $K \to \pi\pi$ amplitudes already incorporate a large number of sophisticated algorithmic and theoretical techniques:
  - G-parity boundary conditions to get two moving pions in a box with the correct kinematics
  - Lanczos algorithm to calculate low modes
  - All-to-all propagators with low mode deflation
  - Pion sources and sinks split in time
  - Subtraction of the vacuum intermediate state to reduce statistical errors
  - Non-perturbative renormalization to normalize the matrix elements correctly
- Faster computers will reduce statistical errors
- The systematic errors can also be markedly reduced with more computation.
- Current steps to improve the calculation:
  - Improve methodology: zMobius formulation for the quarks
  - Ready our software for the next generation of computers:

Improving Domain Wall Fermion Algorithms

- These complicated kaon properties can only be calculated if our discretized system has all of the symmetries of the continuum system.
- Domain wall fermions provide this, by adding a fifth dimension to our problem and localizing the left and right handed quarks on opposite boundaries of the fifth dimension



- This increases the computation cost by substantial factor
- Reducing L_s, the length of the fifth dimension, while preserving the symmetries, reduces the cost.
- Mobius Domain Wall Fermions are a variant formulation with reduced L_s (arXiv:1206.5214)
- Making the parameters in the Mobius formulation complex reduces L_s further. This is called zMobius and was developed by Blum, Izubuchi, Jung, Lehner and Syritsyn at BNL.



- Can reduce L_s from 32 to 14 with very little change in the Dirac operator
- The number of iterations for a conjugate gradient solution increases markedly when moving from the original Domain Wall Fermion formulation to Mobius to zMobius.
- Improved preconditioners ameliorate this somewhat, but further improvements in preconditioners would be useful and would immediately lead to faster solves.
- Coding effort required to implement complex parameters and variant preconditioners in our high-performance codes (Jung).

zMobius in Generation of QCD Ensembles

- For precision measurements of the K_L–K_s mass difference, we need QCD ensembles that include a physical charm quark and have small lattice spacing (1/a ≈ 3 GeV).
- The RBC and UKQCD Collaborations are generating 2+1+1 flavor Mobius DWF ensembles on Mira at the ALCF.
- Greg McGlynn (Columbia) and Chulwoo Jung (BNL) have implemented zMobius into our evolution code.
- We are running in production now on 12 racks of Mira, using zMobius with L_s = 14 in the molecular dynamics part of our L_s= 32 evolution.
- We see a factor of about 1.5 speedup from the use of zMobius.

zMobius in Measurements

- The all mode averaging technique of Blum, Izubuchi and Shintani (PRD 88 (2013) 9, 094503) allows many measurements to be done with reduced precision, or some other approximation. A correction term calculates the deviation between the approximation and the precise value, but is calculated less often, resulting in substantial reduction in cost.
- zMobius can easily play the role of the approximation, pushing to very small values for L_s.
- Currently in use for nucleon and heavy quark measurements. Being extended to other measurements, such as kaon physics.
- Speed-ups by a factor of a few are expected, in addition to O(10-100) speed-up already achieved with all mode averaging and deflation techniques.

Software Preparation for Next Generation Architectures

- The RBC and UKQCD Collaborations have used highly optimized BGQ code written by Peter Boyle of the University of Edinburgh for the conjugate gradient solvers in our ensemble generation and precision kaon calculations
- Peter Boyle has developed a new data parallel QCD library, called Grid, to exploit the next generation of architectures, particularly the Knights Landing from Intel.
- Grid is under active development by Boyle and Yamaguchi at the University of Edinburgh and Cossu at KEK in Japan, with more contributions from BNL and RBC members beginning.
- Evolving the CPS QCD codes to work with Grid is underway at BNL, lead by Chulwoo Jung.
- Our immediate task is to rewrite our $K \to \pi\pi$ measurement code in Grid, to be ready to run on Cori, or other KNL hardware, by early 2016.
- Some information about Grid is presented here, based on Boyle's poster at Lattice 2015

  https://indico2.riken.jp/indico/getFile.py/access?contribId=24&sessionId=15&resId=0&materialId=poster&confId=1805
- Because Grid's underlying data layout allows parallelizatoin at the SIMD, OpenMP and MPI level, very high efficiency is achieved for code that is written in C++.

We present progress on a new C++ data parallel QCD library. It enables the description of cartesian fields of arbitrary tensor mathematical types.

Ddata parallel interface, conformable array syntax with Cshift and masked operation (c.f. QDP++, cmfortran or HPF).

Three distinct forms of parallelism are transparently used underneath the single simple interface:

- **MPI task parallelism**
- **OpenMP thread parallelism**
- **SIMD vector parallelism.**

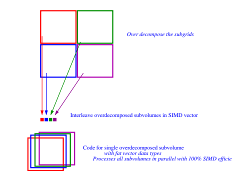
The SIMD vector parallelism achieves nearly 100% SIMD efficiency due to the adoption of a virtual node layout transformation, similar to those in the Connection Machine.

This ensures identical and independent work lies in adjacent SIMD lanes. SSE, AVX, AVX2, AVX512 and Arm Neon SIMD targets are implemented.

The library is under development. Solvers for Wilson, Domain, and multiple 5d chiral fermions (Cayley, Continued fraction, partial fraction) are implemented.

GRID parallel library

- Geometrically decompose cartesian arrays across nodes (MPI)
- Subdivide node volume into smaller *virtual nodes*
- Spread virtual nodes across SIMD lanes
- Use OpenMP+MPI+SIMD to process conformable array operations
- Same instructions executed on many nodes, each node operates on four virtual nodes



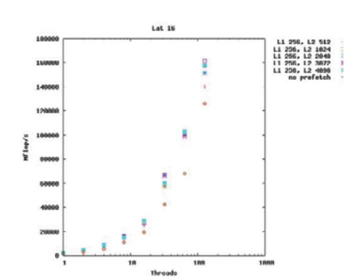
- Conclusion: Modify data layout to align data parallel operations to SIMD hardware
- Conformable array operations simple and vectorise perfectly

```
std:vector<int> grid ({ 8,8,8,8 });
std:vector<int> simd_layout ({ 1,1,2,2 });
std:vector<int> mpi_layout  ({ 1,1,1,1 });

CartesianGrid Grid(grid,simd_layout,mpi_layout);

LatticeColourMatrix A(Grid);
LatticeColourMatrix B(Grid);
LatticeColourMatrix C(Grid);

A = B * C;
```



SU3xSU3 XeonPhi

- New algorithms and evolution of our code base will allow us to exploit the new resources as soon as they become available.
- Precision calculations continue to improve constraints on the Standard Model

Higher resolution increase accuracy and exposes multi-scale physics.

The “turbulent” vacuum and the huge mass scales of the quarks has profound effects on the physics and the need for complex algorithms and codes.

Lattice field theory is *no longer* a simple software problem!

In the last couple of years, advances in Hardware, Algorithm/Software (thanks to SciDAC!) now allow full access to hadronic physics

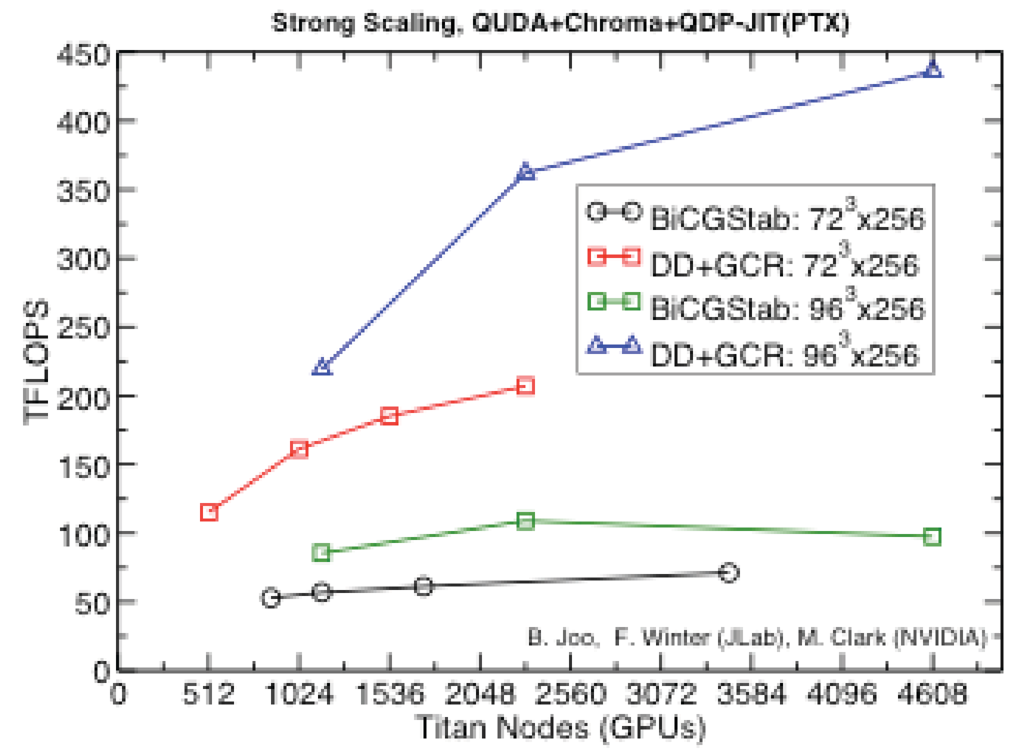
$$a(lattice) \ll 1/M_{proton} \ll 1/m_\pi \ll L(box)$$
$$0.06 \text{ fermi} \ll 0.2 \text{ fermi} \ll 1.4 \text{ fermi} \ll 6.0 \text{ fermi}$$

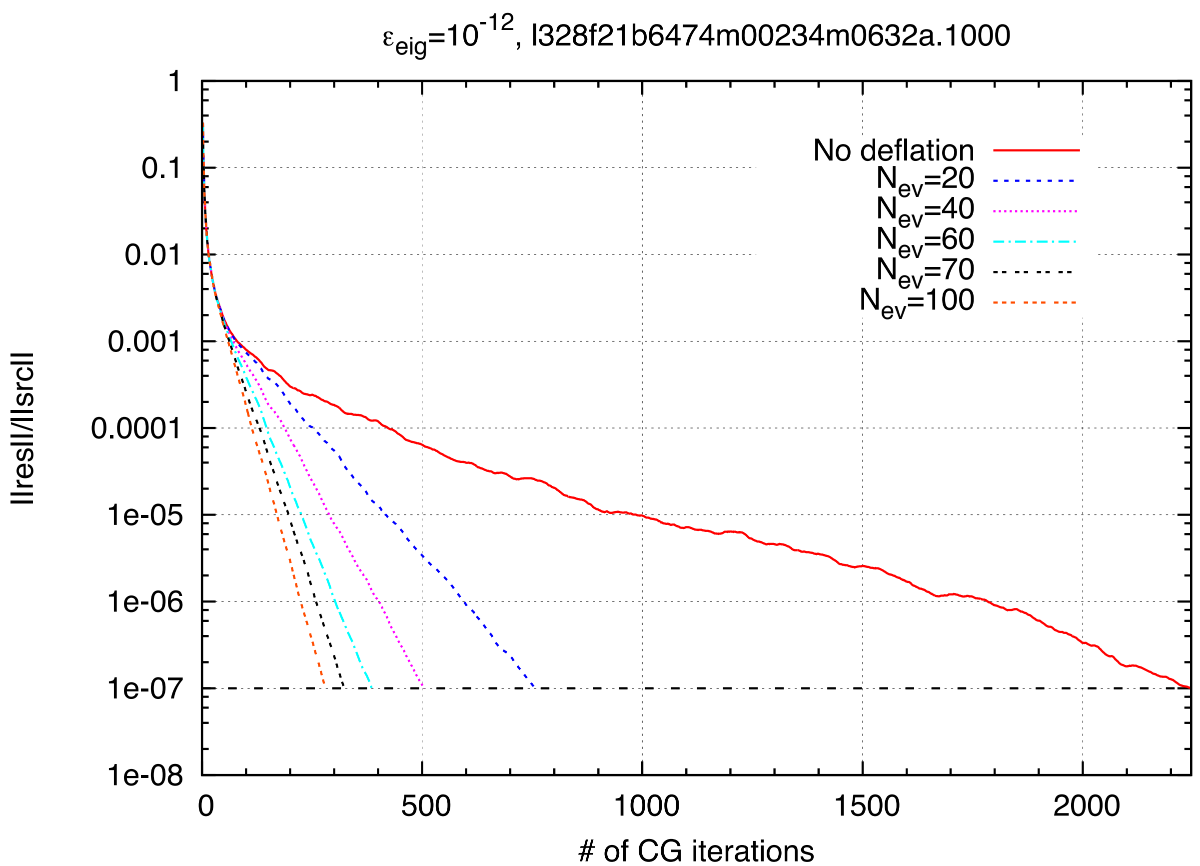
with spacetime lattices on order L^4 = (100)^4 and larger.

The numerical results combined with a growing arsenal of theoretical tools (heavy quark effective expansions, exact chiral expansion, RG scaling to zero lattice spacing (a =0) and infinite volume(1/L =0) limits give some high precision QCD predictions critical to the HEP/NP experimental program. This will become increasing important in the coming decade.

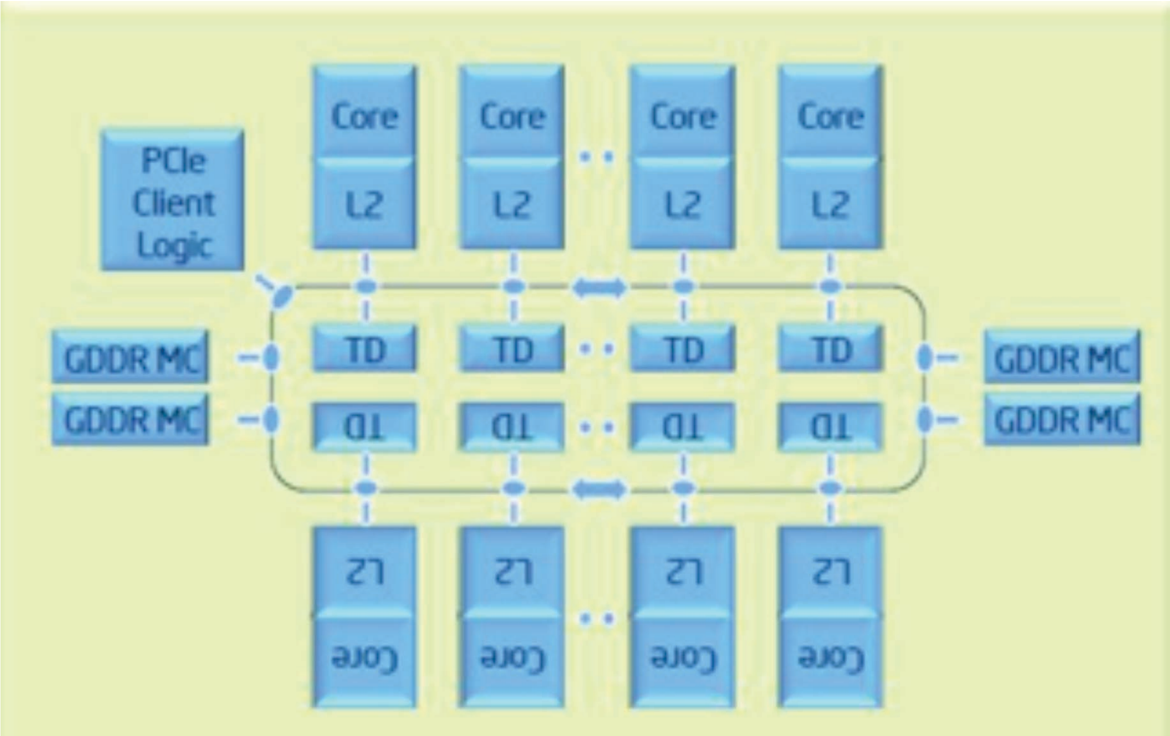
Domain Decomposition & Deflation

- 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





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

Short List of Current HEP Software Priorities

1. Finish Optimizing Multigrid for Wilson Clover on GPUs (FNAL/NVIDIA/BU) & integrate into Chroma of NP
2. Staggered Multigrid on GPUs and Portable C code(FNAL/NVIDIA / BU)
3. HMC evolution optimization using multi-scale methods for Wilson (started) and Domain Wall (early feasibility study at BNL)
4. Portable framework development to target both GPU and PHI systems* FUEL (at Argonne) and GRID (at Edinburgh/BNL)
5. On going code development for MILC (including Utah/Illinois/Arizona) and CPS (BNL/RBC collaboration)
6. Co-ordination with NP projects described by Balint Joo yesterday. Specifically strong scaling improvements with communication mitigation domain decomposition mixed precision etc.

* Critical resource: Our close collaboration with software engineers at NVIDIA and INTEL and early access to hardware at Oak Ridge, Argonne and NERSC as well as clusters at Jlab/FNAL/BNL.

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
- MIT
  - Andrew Pochinsky
- Boston University
  - Rich Brower
  - Claudio Rebbi
  - Mike Clark
  - James Osborn
  - Saul Cohen
- Michael Cheng
- Oliver Witzel
- Penn State
  - James Brannick
  - Ludmil Zikatanov
- Tufts
  - Scott MacLachlan
- Argonne
  - James Osborn
- NVIDIA
  - Mike Clark
  - Ron Babich
- INT Seattle
  - Saul Cohen