Lattice QCD on GPUs using Chroma and QUDA

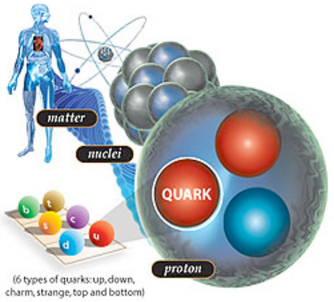
Bálint Joó, Frank Winter
US DOE Jefferson Lab

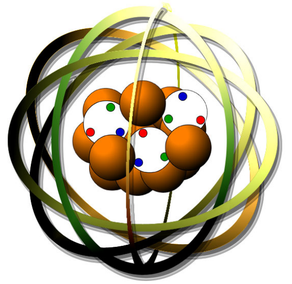
Mike Clark, NVIDIA

NVIDIA GPU Technology Conference
San Jose, CA
March 20, 2013

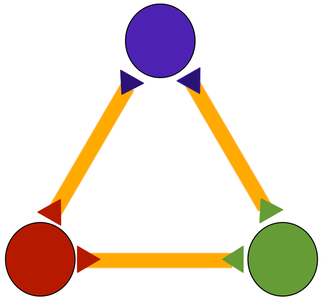
Quantum Chromodynamics

- Quantum Chromodynamics is the theory of the nuclear strong force
- Matter is made up of atoms
- The nuclei of atoms are made up of protons and neutrons
- The protons and neutrons are made up of quarks and gluons
- Quarks gluons carry so called “color” charges.
- Only “colorless” combinations can be seen in nature

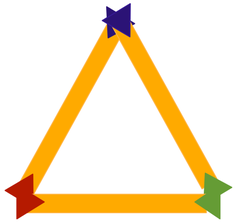




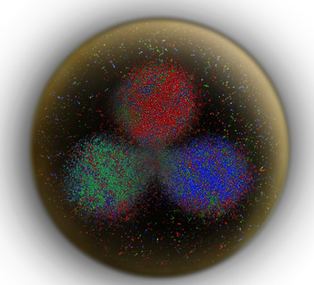
meson: 2 quarks



baryon: 3 quarks

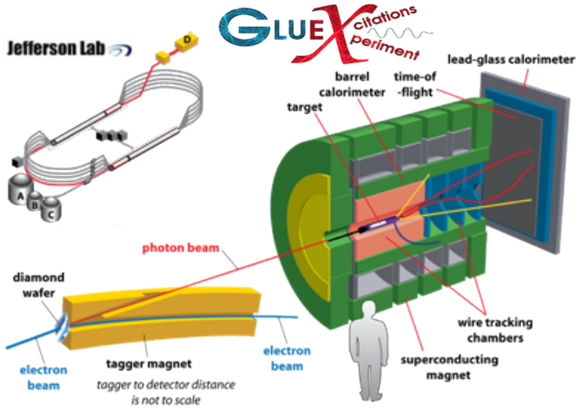


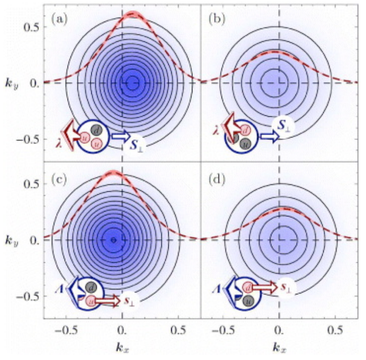
glueball: 0 quarks
only gluons



Important questions in Nuclear Physics

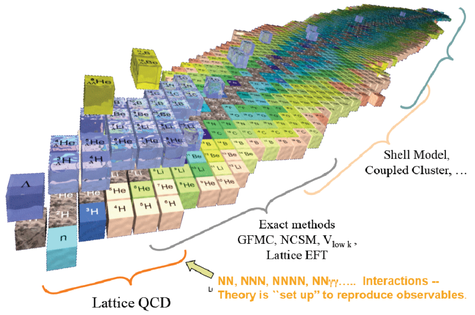
- What observable states does QCD allow?
  - what is the role of the gluons?
  - what about exotic matter?





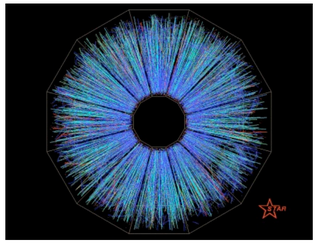
Hägler, Musch, Negele, Schäfer, EPL 88 61001

- How does QCD make protons, neutrons?
  - what are the distribution of quarks, gluons, etc in a proton or neutron ?



- QCD must predict properties of light nuclei
  - how to make helium, tritium etc

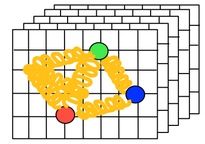
- How does QCD behave under extreme temperatures & pressures such as in supernovae or shortly after the Big-Bang.



LQCD Calculation Workflow

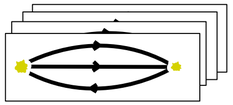
Gauge Generation

Gauge Configurations

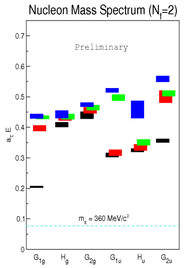


Analysis Phase 1

Propagators, Correlation Functions



Analysis Phase 2

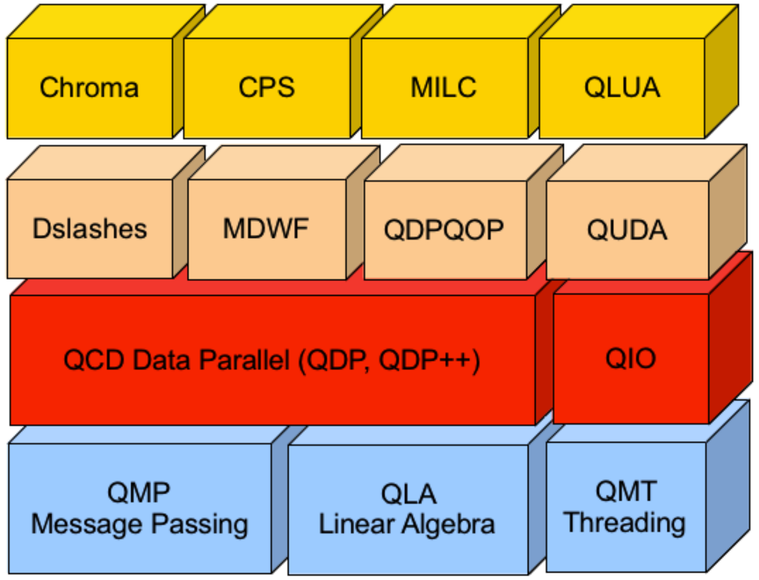


Physics Result

- Gauge Generation: Capability Computing on Leadership Facilities
  - configurations generated in sequence using Markov Chain Monte Carlo technique
  - focus the power of leadership computing onto single task exploiting data parallelism
  - strong scaling challenge.
- Analysis Phase 1: Capacity computing, cost effective on Clusters
  - task parallelize over gauge configurations in addition to data parallelism

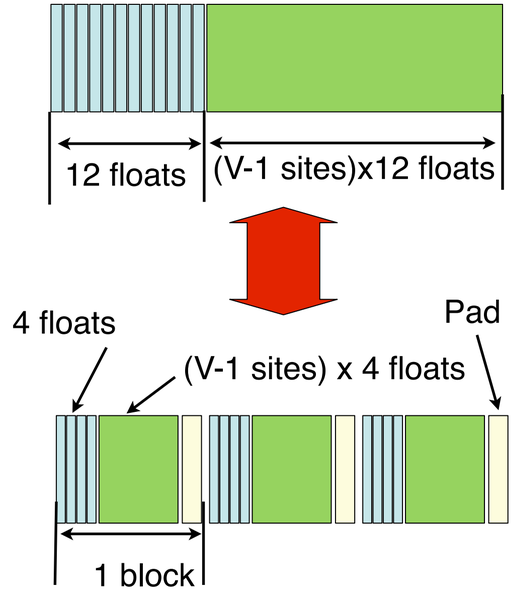
Chroma, QUDA and other software

- Chroma is an application suite for LQCD calculations
  - developed under US DOE SciDAC-1 and SciDAC-2 initiatives
  - facilitates gauge generation and analysis
  - large worldwide user base
- QUDA is a library of LQCD components for NVIDIA GPUs
  - provides optimized solvers, some force-terms
- Chroma and QUDA have a synergistic partnership
  - QUDA enabled Chroma on GPUs
  - Chroma wrapped QUDA and brought it to its large user base



QUDA Performance Optimizations

- LQCD is typically memory bound
  - Dslash: Nearest neighbour stencil in 4D
    - Wilson quarks: 0.92 FLOP/B (SP)
    - Staggered quarks: ~0.66 FLOP/B (SP)
- Lay out data for coalesced memory access
- Use symmetries to compress SU(3) matrices
  - 2 row storage or 8 parameter storage
  - reconstruct 3rd row with “free” FLOPs
- Use reduced precision if possible (e.g. 16bit)
  - mixed precision solver , iterative refinement + reliable updates
- Fuse BLAS like kernels - increase reuse



$$\begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ \mathrm{x} & \mathrm{x} & \mathrm{x} \end{pmatrix}$$

$$\begin{aligned} \mathbf{a} &= (a_1, a_2, a_3) \\ \mathbf{b} &= (b_1, b_2, b_3) \\ \mathbf{c} &= (\mathbf{a} \times \mathbf{b})^* \end{aligned}$$

$$\begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{pmatrix}$$

USQCD GPU Clusters

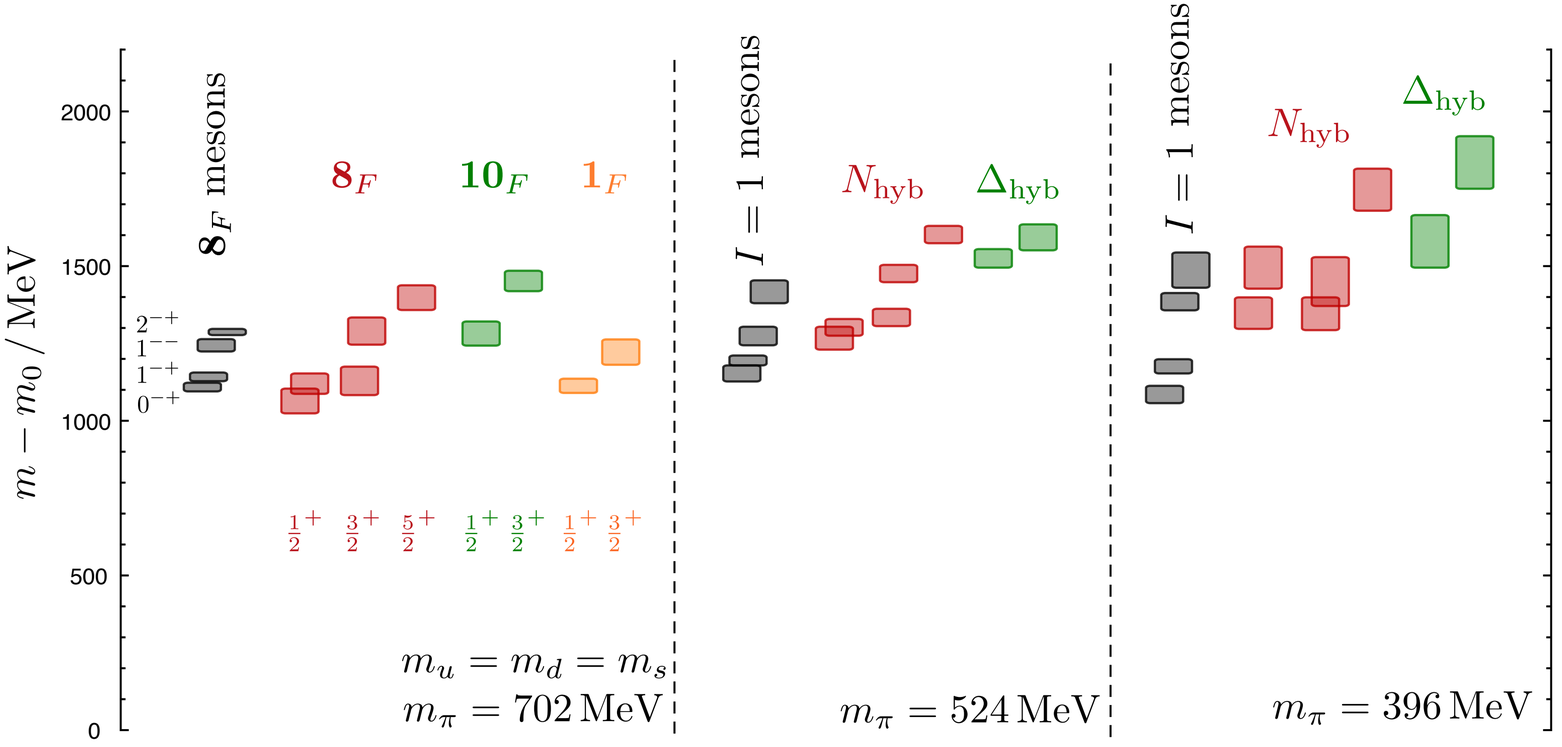
- USQCD National Facility (JLab, FNAL, BNL)
  - JLab and FNAL Operate GPU Accelerated Clusters
  - JLab: 138 quad GPU nodes
    - 41x4 Tesla K20M
    - 34x4 Tesla C2050/M2050,
    - 56x4 GTX 480/580
    - 7x4 GTX 285
  - FNAL: 72 dual nodes (Tesla M2050 GPUs)

JLab 9G GPU Cluster

JLab 10G GPU Cluster

Science From GPU Clusters

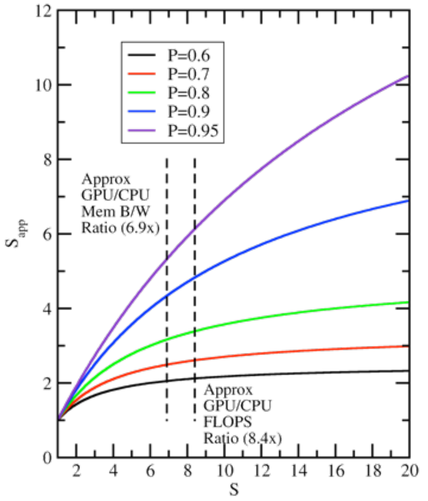
- Hybrid Excitations in mesons, and baryons at a common scale of ~1200 MeV
- Pattern suggests chromo-magnetic excitation
- common in mesons, baryons.
- “Effective degree of freedom” ?
- first principle calculation can agree with or disfavor effective models

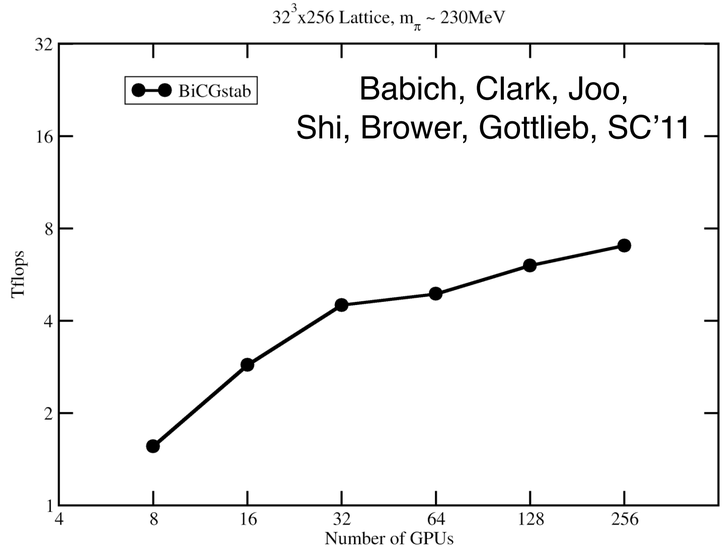


J. J. Dudek, R. G. Edwards, “Hybrid Baryons in QCD”, Phys. Rev. D85, 054016

Challenges in Gauge Generation

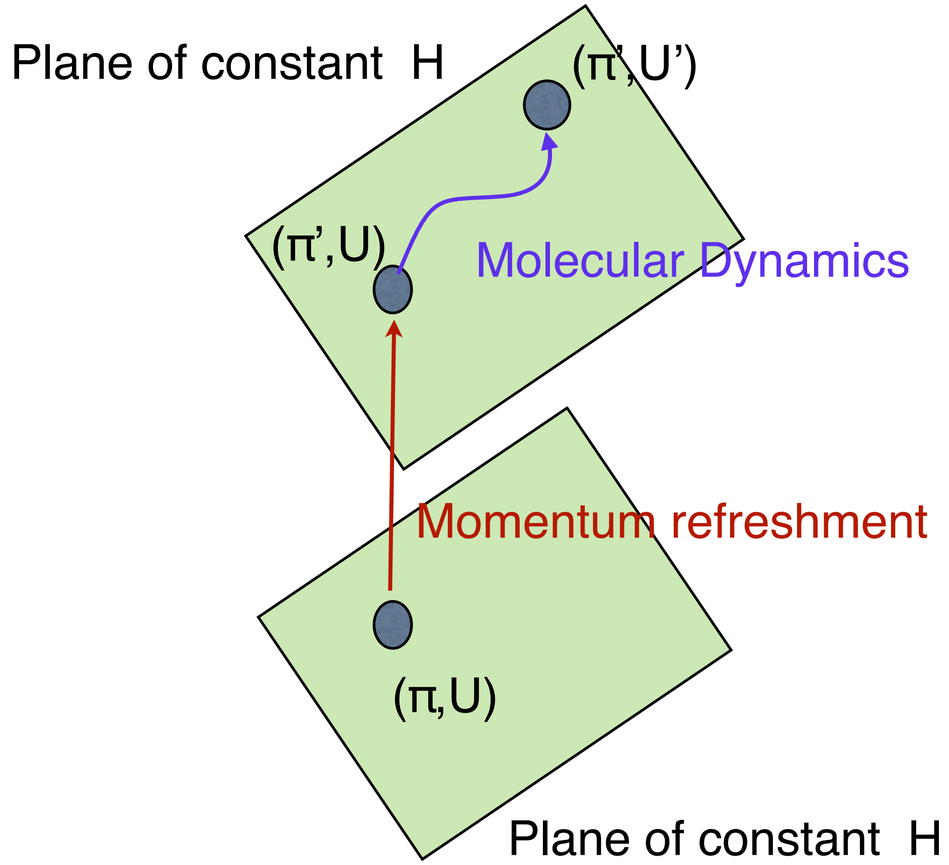
- Amdahl's law effects
  - Unaccelerated code can drag down performance of accelerated code
  - Solution: move more code to GPU
  - Problem: how to preserve 10 year investment in Chroma
  - Solution: target QDP++ layer on which most of Chroma is built
    - QDP-JIT: Frank Winter's talk at this conference
- Strong Scaling
  - As node count increases, local problem size decreases
  - device occupancy is reduced, surface to volume ratio increases
  - latencies start to become important
  - Solution: hardware and software improvements





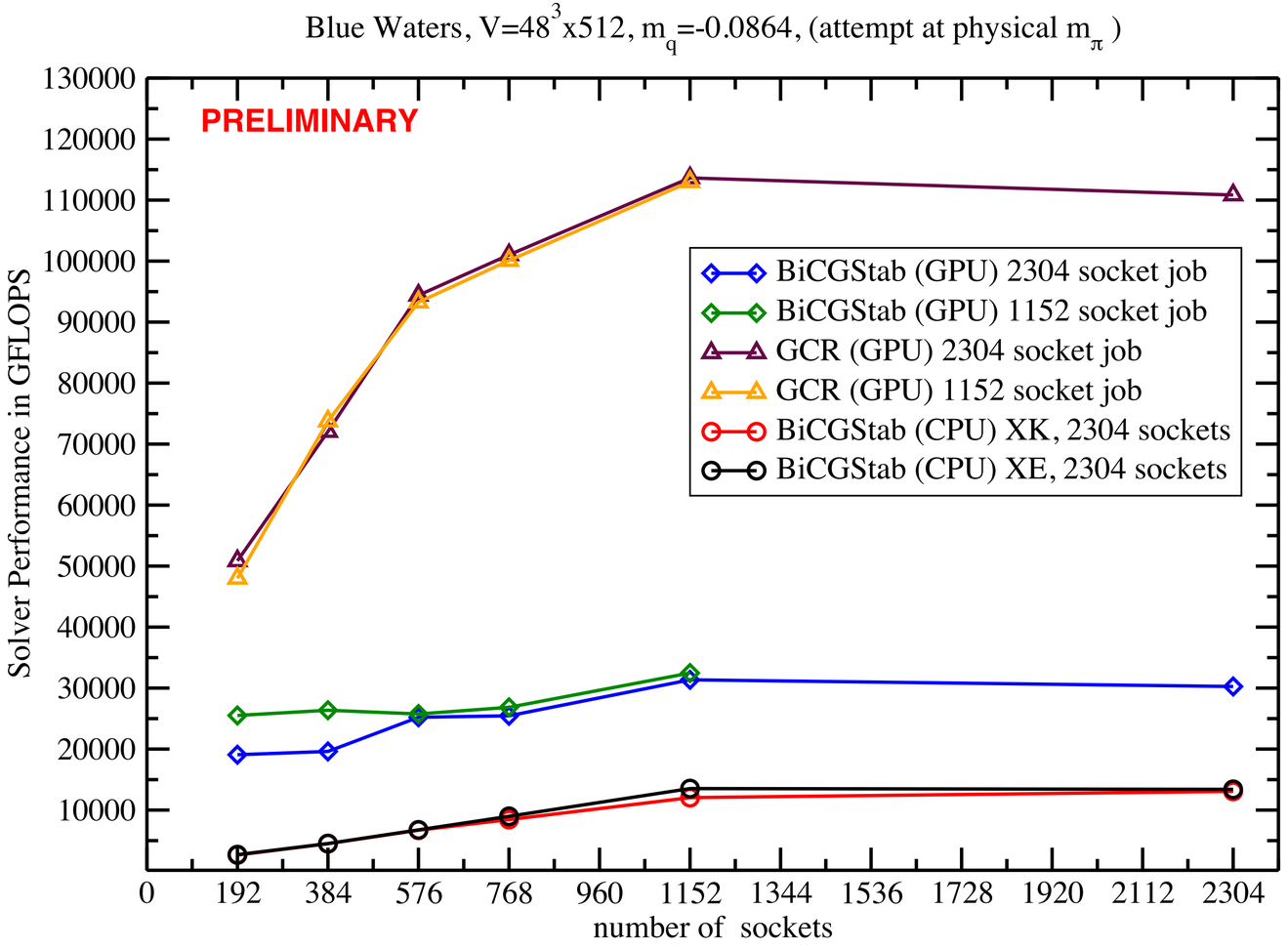
Gauge Generation

- Hybrid Molecular Dynamics Monte Carlo
  - new state from old using Molecular Dynamics (MD)
  - Metropolis Accept/Reject Step
- Typically >90% of time spent in MD Forces
- Force term for quarks has 3 main components
  - Solution of the Dirac Equation
  - Derivative of the Fermion Matrix
  - Optional: Derivative of smeared links w.r.t thin links
  - Last two of these provide Amdahl's law slowdown
- Non Quark Forces provide Amdahl's law slowdown



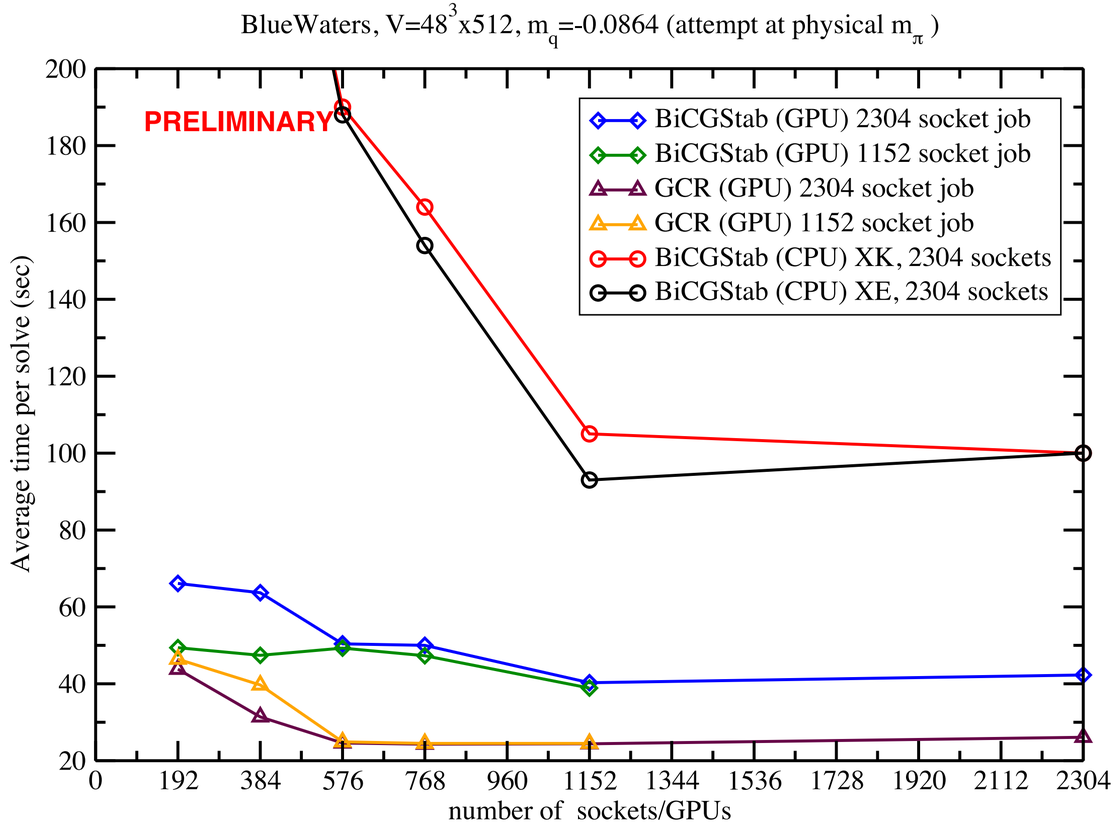
QUDA Solver Scaling: FLOPS

- 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

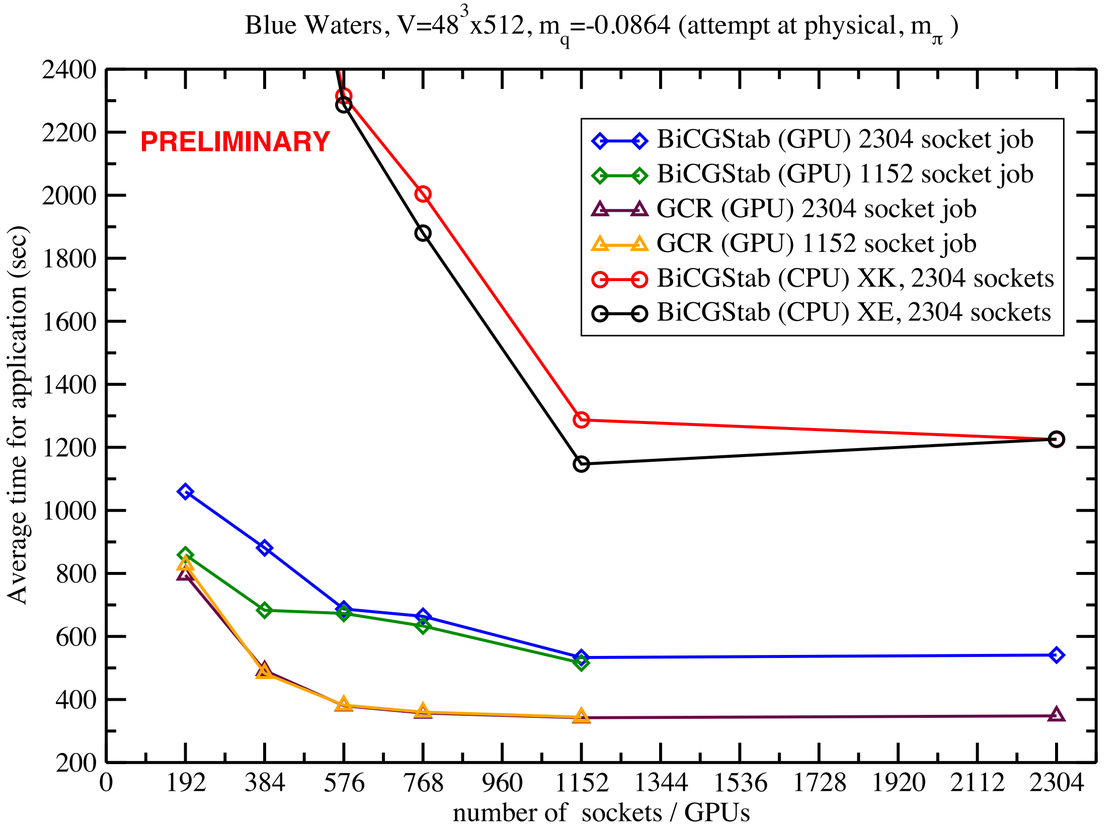


QUDA Solver Scaling: Wallclock Time

Time per solve

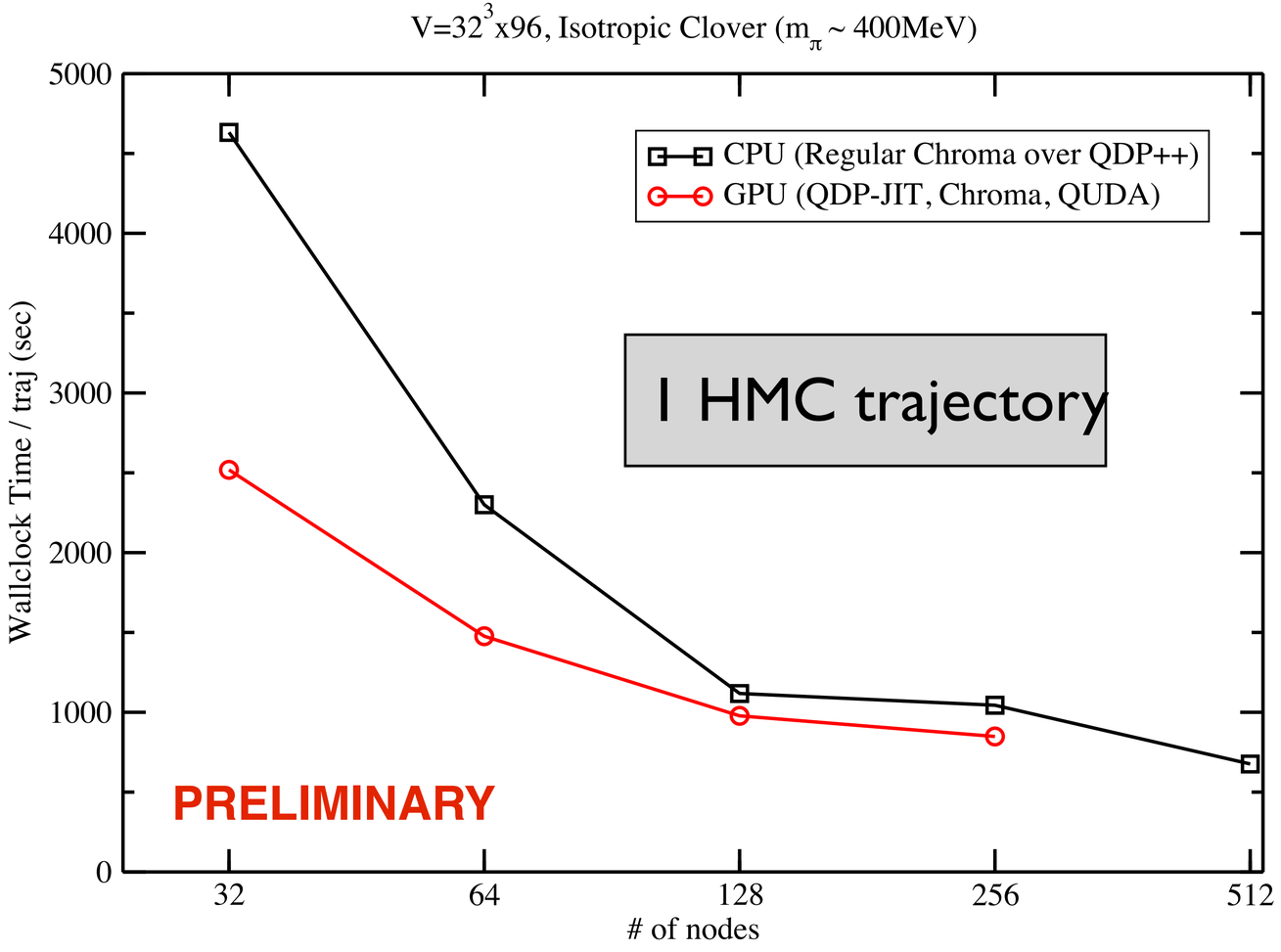


Whole application time



Gauge Generation using QDP-JIT/C

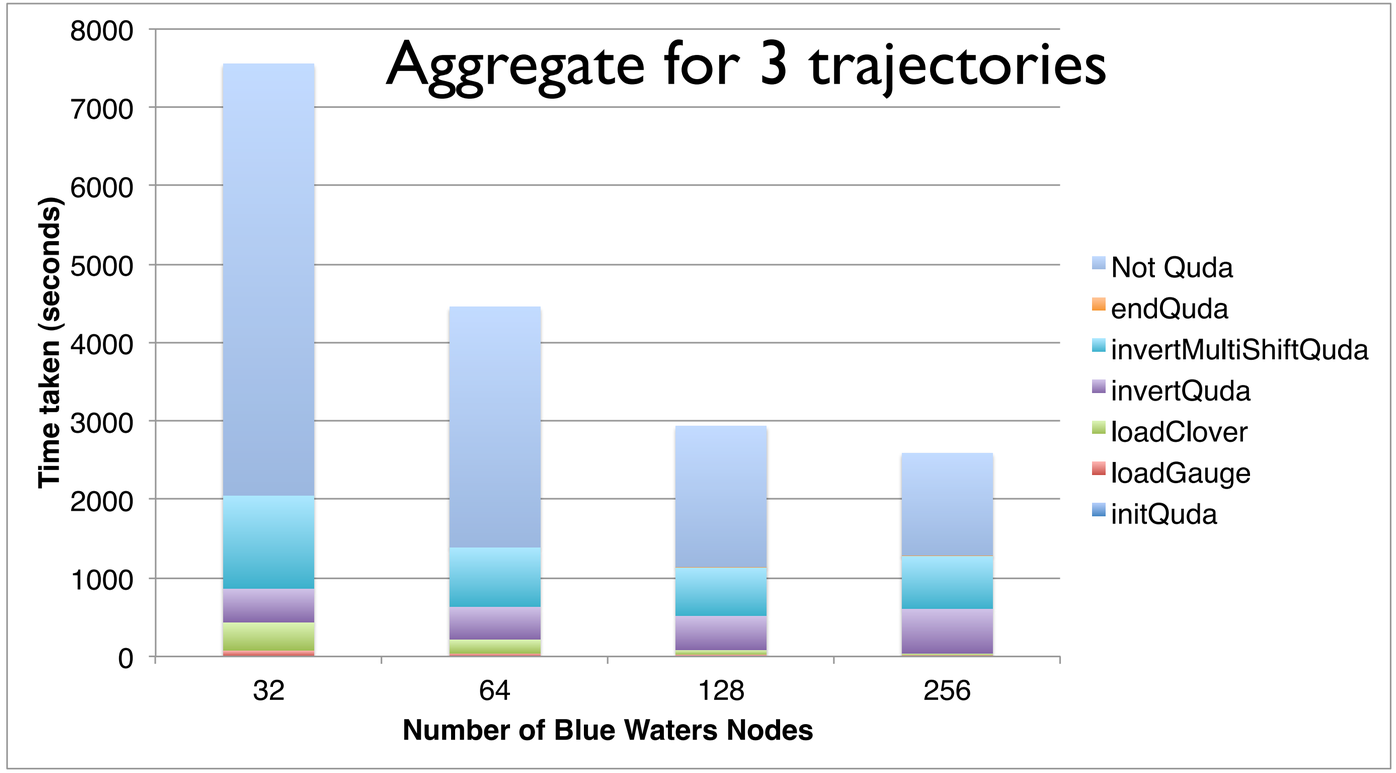
- Initial result: 2+1 flavor clover gauge generation on Blue Waters
- Full featured:
  - Full clover action
  - stout smearing of the links
- Solvers from QUDA
  - BiCGStab for 2 flavor piece
  - multi-shift CG for 1 flavor piece
- Non solver part through QDP-JIT/C
  - JIT to CUDA/C - see Frank Winter's talk
- This is a milestone for us !!!



Gauge Generation Using QDP-JIT/C

- NB: 32^3x96 lattice is very small
  - Severe strong scaling challenge
- Solver scaling has ‘topped out’
  - GCR in HMC is work in progress.
  - Optimize communications further in QUDA?
- Amdahl’s law effects still substantial
  - QDP-JIT/PTX will help
  - May need more hand tuned kernels outside of the solvers
- Expect 40^3x256 lattice to scale to ~1000-2000 nodes with GCR solver.

PRELIMINARY



Summary and Outlook

- Chroma and QUDA have a synergistic, mutually beneficial relationship
- The JIT technology allows all of Chroma (anything written in QDP++) to run on GPU.
- Milestone: Our Gauge Generation program can run on GPUs, Blue Waters, Titan
- Milestone: Sufficiently large configuration scaled to over 1000 GPUs
- Prospects & Challenges:
  - PTX version of QDP-JIT will improve remaining Amdahl's law effects
  - Extreme strong scaling is still challenging
  - Improvements to come:
    - Algorithmic: more scalable solvers in QUDA
    - Lower level: better exploitation of hardware comms features (GPU Direct etc).
    - Potential hardware improvements in future systems (beyond PCIe2)

More Lattice at GTC

- Talks:
  - Mathias Wagner: “GPUs Immediately Relating Lattice QCD to Collider Experiments”*
  - Frank Winter: “QCD Data Parallel (Expressive C++ API fr Lattice Field Theory) on GPUs *
  - Hyung-Jin Kim: “Columbia Physics System with QUDA”, Thurs. 4:30pm, Room 111
- Posters:
  - Hyung-Jin Kim: “CPS with QUDA”
  - Richard Forster, Ágnes Fülöp: “Yang-Mills Lattice on CUDA”
  - Alexei Strelchenko: “Extending the QUDA library with Twisted Mass Fermions”

- * = these talks preceded mine, but you should be able to find recordings of them to check out.

Acknowledgements

- Computer Time:
  - NSF Blue Waters Facility at NCSA,
  - Titan (and TitanDev) Facility at Oak Ridge Leadership Computing Facility (OLCF),
  - US National Computational Facility for Lattice Gauge Theory (Jefferson Lab Clusters)
- Funding:
  - Partial support for this work was provided through Scientific Discovery through Advanced Computing (SciDAC) program funded by U.S. Department of Energy, Office of Science, Offices of Advanced Scientific Computing Research, Nuclear Physics and High Energy Physics.