LQCD SciDAC-5

Robert Edwards
Jefferson Lab

A few notes

- CEBAF is running; users are collecting data
- Please wear your name badge at all times
  - Receptionists in the buildings will know to let you in

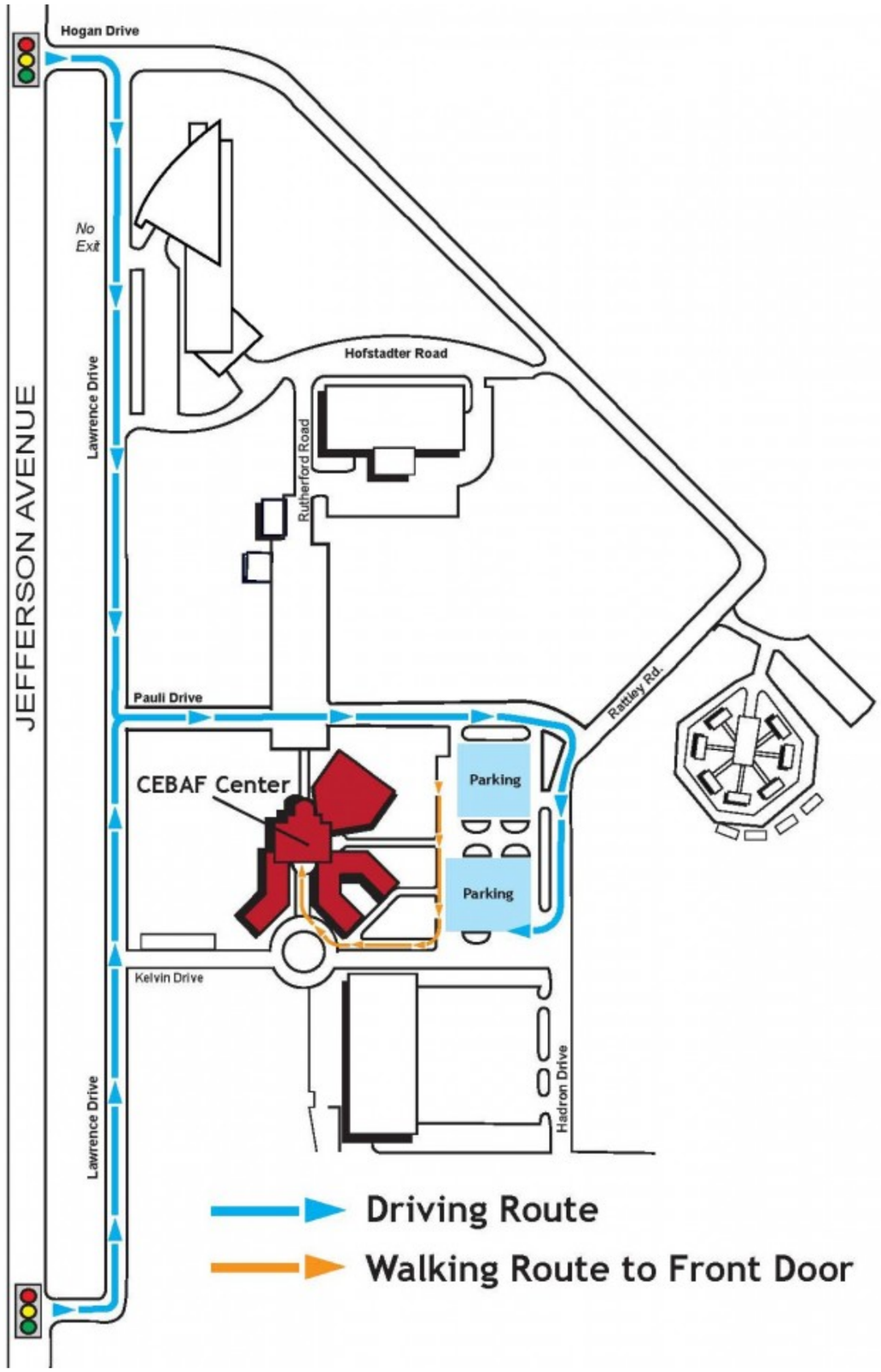
- This is a public event at a DOE facility
  - Please stay in the public areas of the campus
  - *All information to be presented or discussed must meet DOE standards for public release and no information will be presented or discussed that is proprietary or protected from release by statue, regulation, DOE policy.*

The surroundings

Shops in Tech Center

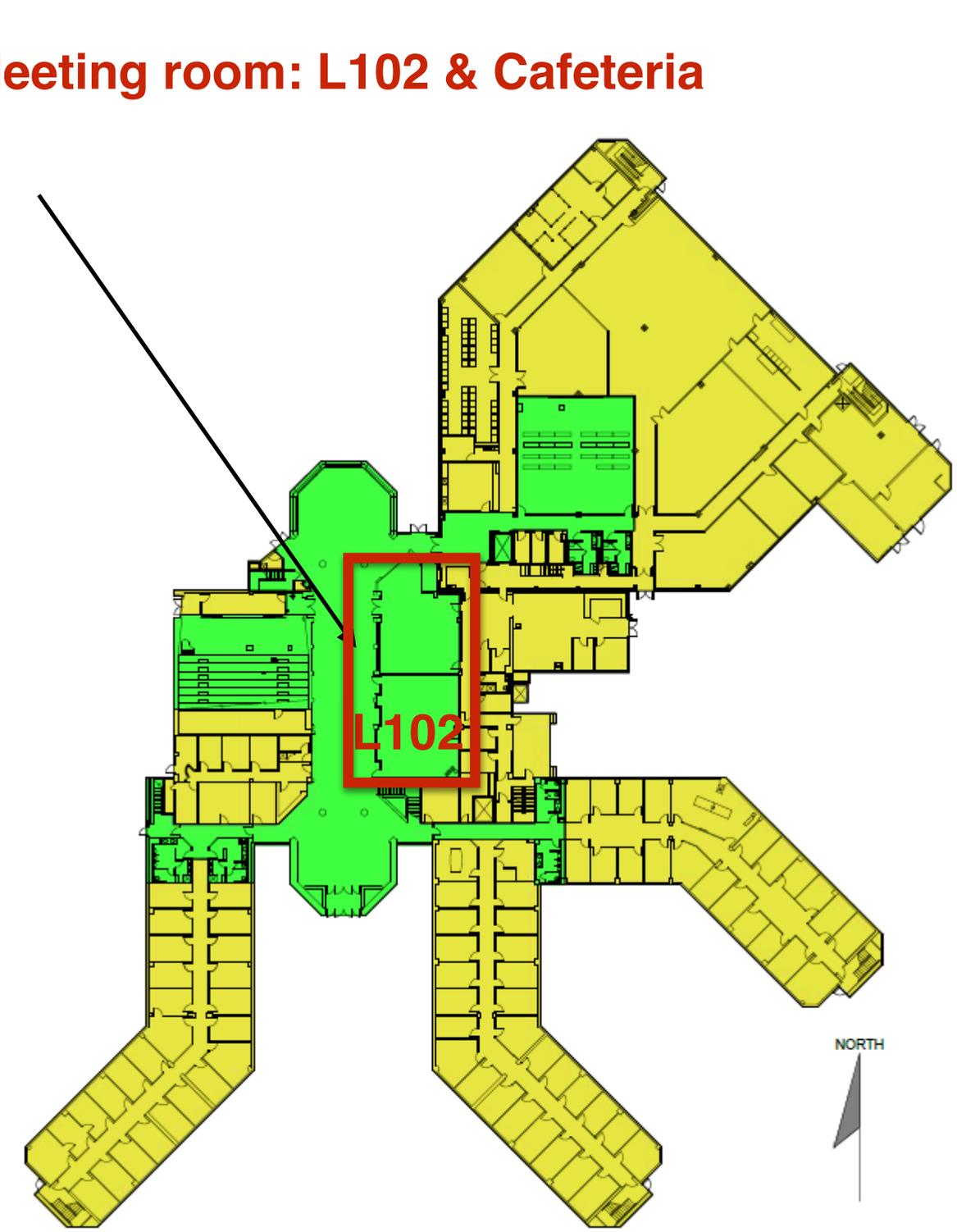
Breakfast+Lunch:
JLab Cafeteria

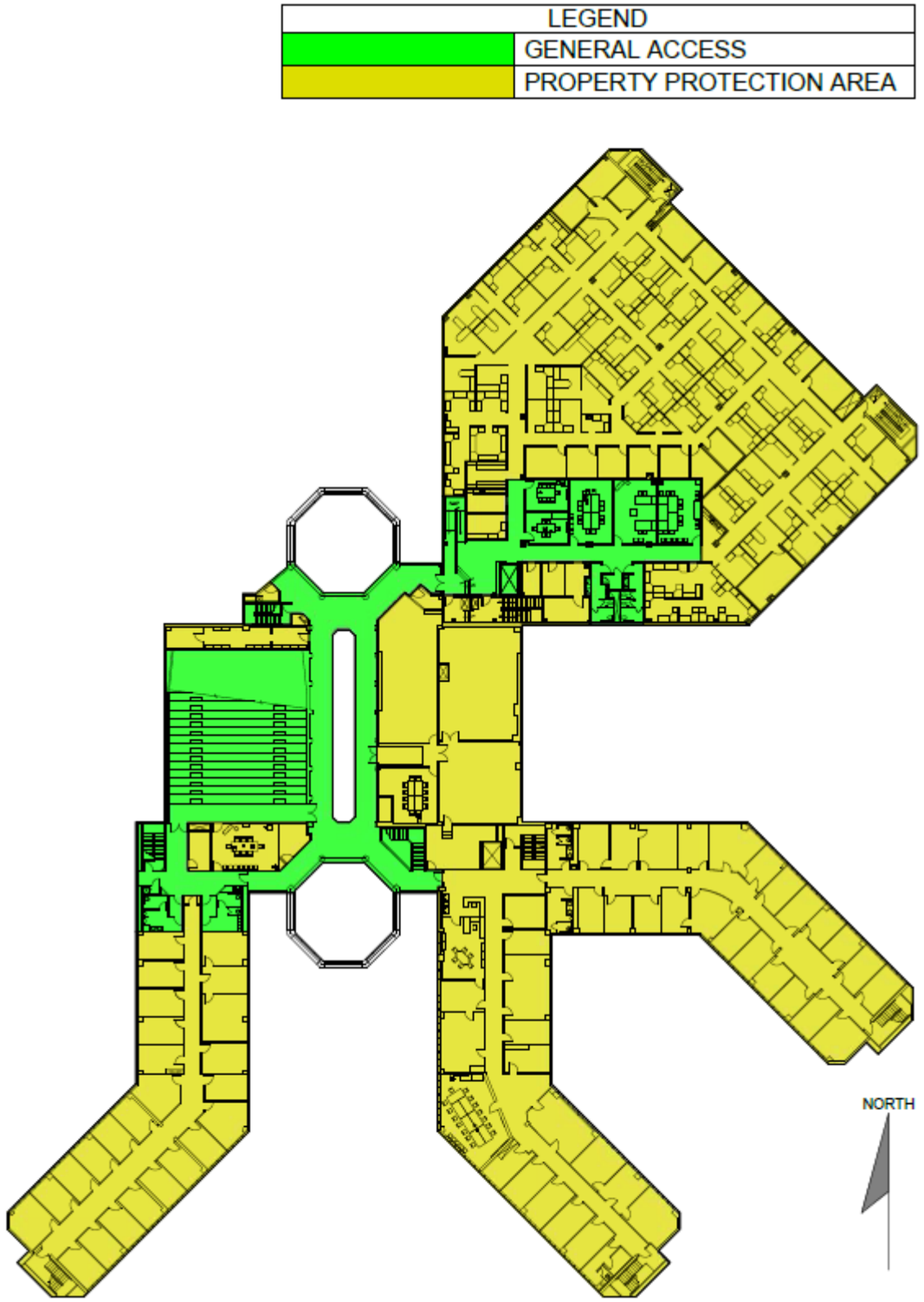
CEBAF Center @ JLab



CC meeting locations

Meeting room: L102 & Cafeteria

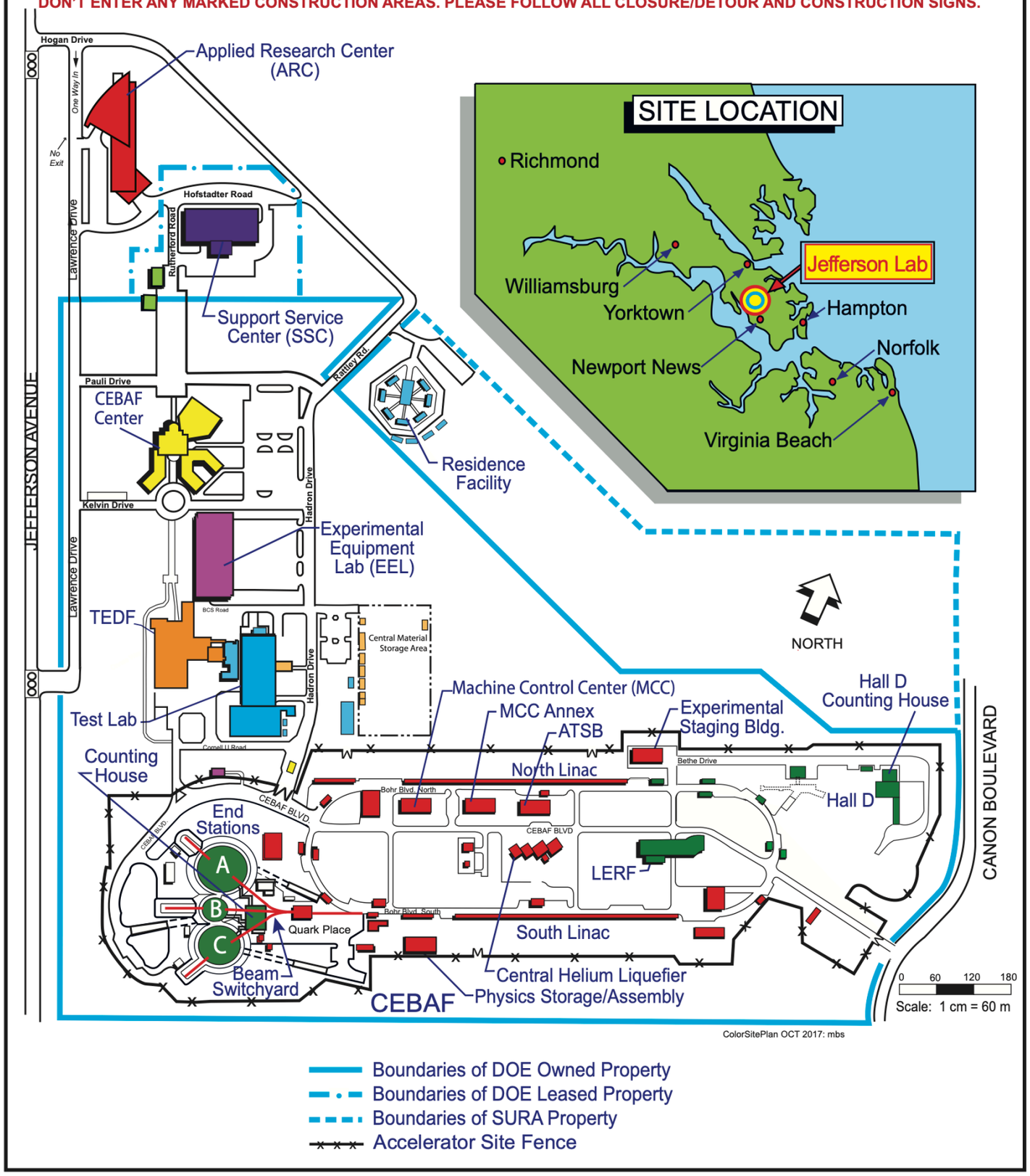




SITE PLAN

Jefferson Lab

USE CAUTION WHILE DRIVING AND/OR WALKING AROUND JLAB. SEVERAL CONSTRUCTION PROJECTS ARE UNDERWAY. DON'T ENTER ANY MARKED CONSTRUCTION AREAS. PLEASE FOLLOW ALL CLOSURE/DETOUR AND CONSTRUCTION SIGNS.



LQCD ASCR/NP SciDAC-5

Fundamental nuclear physics at the exascale and beyond

- Argonne:
  - Prasanna Balaprakash
  - Robert Latham
  - Yong Zhao
- Brookhaven:
  - Swagato Mukherjee
- JLab
  - Jie Chen
  - Robert Edwards
  - Eloy Romero
  - Frank Winter
- Lawrence Berkeley
  - Aydin Buluc (& Oguz Selvitopi)
  - Sherry Li
- Los Alamos
  - Boram Yoon ->
  - Tanmoy Bhattacharya (& Jun-Sik Yo)
- MIT:
  - Saman Amarasinghe
  - Will Detmold
  - Andrew Pochinsky
  - Phiala Shanahan
- Oak Ridge:
  - Balint Joo
- William & Mary:
  - Kostas Orginos
  - Andreas Stathopoulos

Publicity and useful repositories

GitHub: https://github.com/LQCDSciDAC Website: https://lqcdscida.github.io

ASCR/NP LQCD SciDAC Project

Home About SciDAC Highlights People Publications

ASCR/NP LQCD SciDAC Project

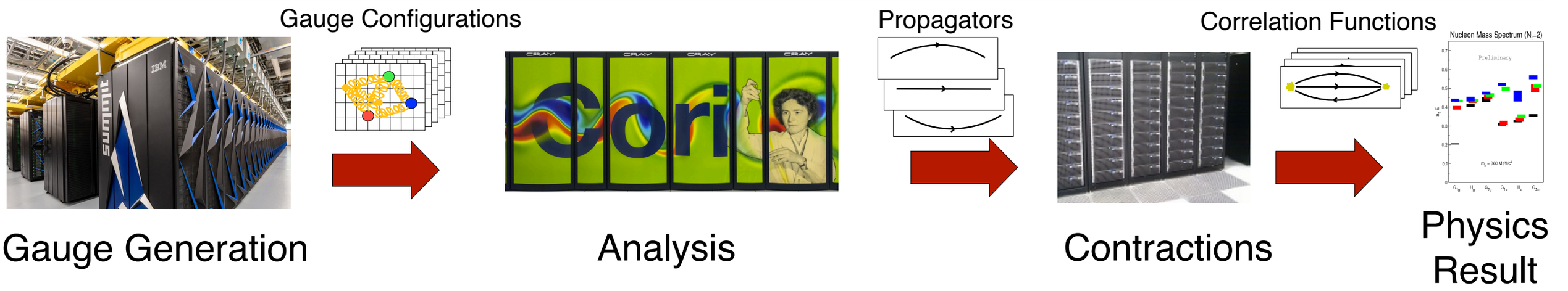
The Lattice Quantum Chromo-Dynamics (LQCD) ASCR/NP SciDAC Project is supported by the U.S. Dept. of Energy Office of Nuclear Physics and the Office of Advanced Scientific Computing Research. This SciDAC project focuses on an ambitious program of theoretical, algorithmic and software development which will enable calculations using lattice Quantum Chromodynamics (LQCD) methods to exploit the new generation of leadership-class resources and dedicated hardware to address fundamental questions in nuclear science. Specifically, our project will impact our understanding of results from current heavy ion experiments at the Relativistic Heavy-Ion Collider (RHIC), the study of excited and exotic states of hadrons at CLAS-12 and GlueX at Jefferson Lab (JLab) and the hadron and nuclear structure programs at RHIC-spin and JLab. The calculations that are enabled by the proposed developments will also look forward to experiments on protons and nuclei at the upcoming Electron-Ion Collider (EIC).

Organized around science goals

Science areas recognized in the 2015 NSAC Long Range Plan

- Hadron spectroscopy
  - CLAS12 & GlueX @ JLab
- Hadron structure
  - JLab & future EIC @ BNL
- Partonic structure of nuclei
  - JLab & Future EIC, future DUNE (FNAL) & HyperK (Japan)
- Extreme matter - quark gluon plasma
  - RHIC @ BNL and CERN

Lattice QCD Calculation Workflow



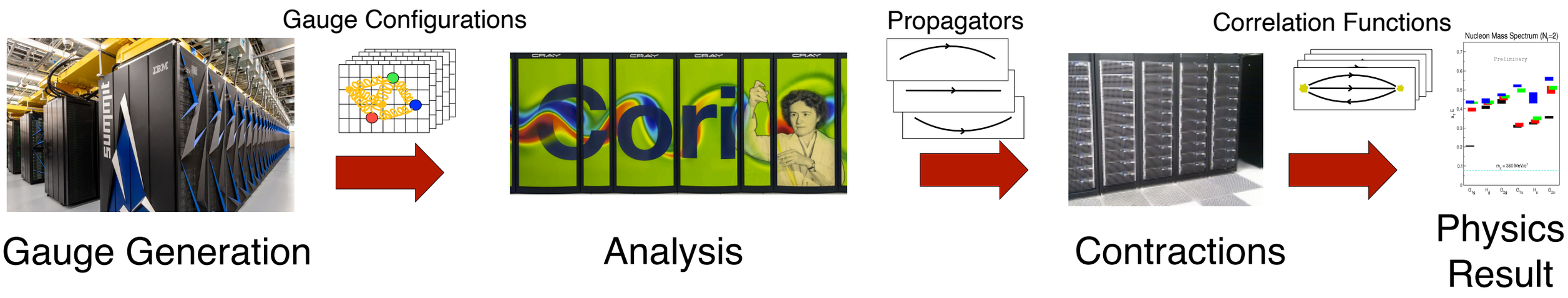
- Core computational unit is solution of Dirac equation - large sparse linear system

- **Gauge Generation**: Few independent chains. Strong scaling challenge
  - Carried out on leadership computing facilities

- **Analysis**: Propagator Calculations
  - Many independent solves, throughput challenge

- **Contractions**: Correlation Function Construction
  - Most cost effective on capacity/midrange systems, tho LCFs can also be used

Challenges in LQCD workflow

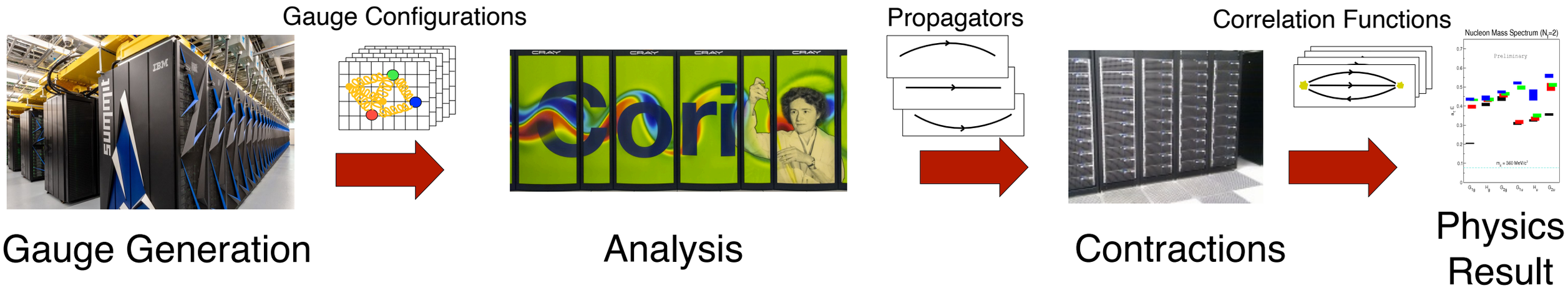


- **Gauge Generation:**
  - Critical slowing down (autocorrelations increase as lattice spacing decreases)

- **Analysis:**
  - Require much higher statistical precision

- **Contractions:**
  - Cost/Complexity rapidly increases for many-body systems

Project development directions

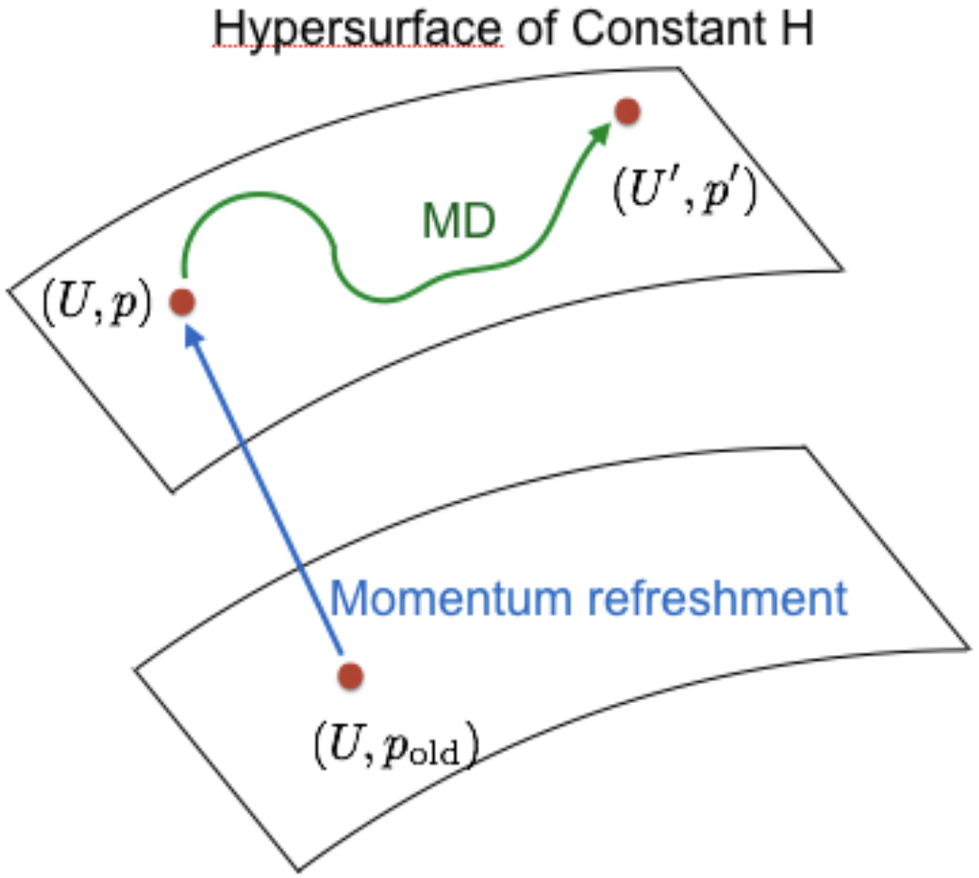


- **Gauge Generation:**
  - Critical slowing down (autocorrelations increase as lattice spacing decreases)
    - Machine-learning based Markov chain updating
- **Analysis:**
  - Require much higher statistical precision
    - Hierarchial integration - combine gauge and analysis
- **Contractions:**
  - Cost/Complexity rapidly increases for many-body systems
    - Improved: graph contractions + factorization + sparse representations
- Improved software support & application readiness

How to produce gauge fields?

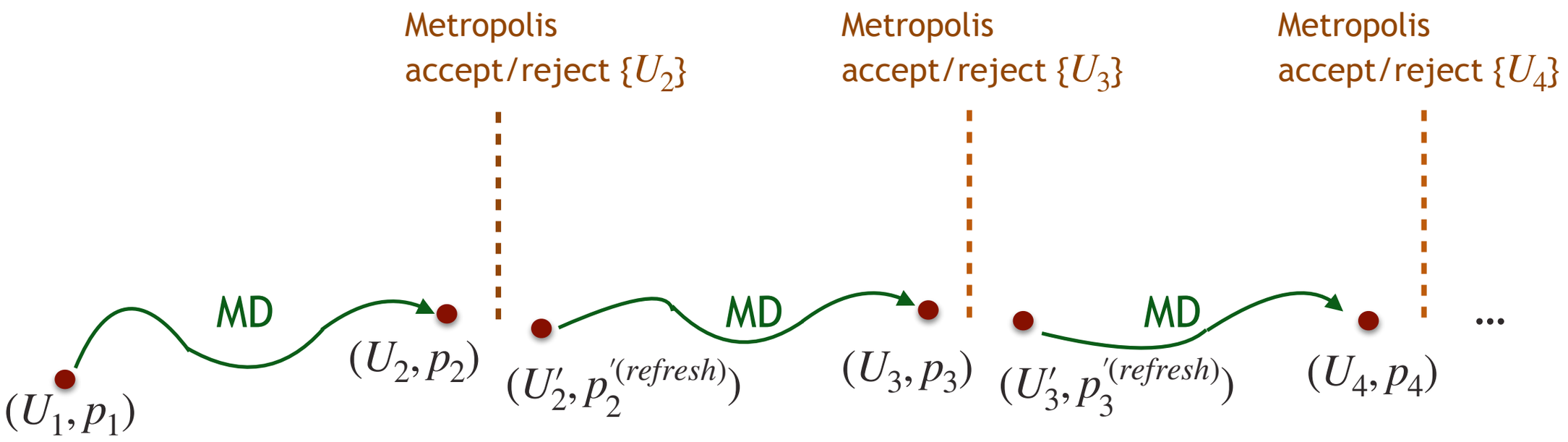
- Hamilton's equations - 1st order coupled differential eqns.
- Each integration step: sparse-matrix solve of Dirac eq. (small # RHS)
- Strong scaling challenge

- Limitations
  - Must be "reversible"
  - No adaptive time steps
  - Symplectic integrators



Metropolis updating via Hybrid Monte Carlo

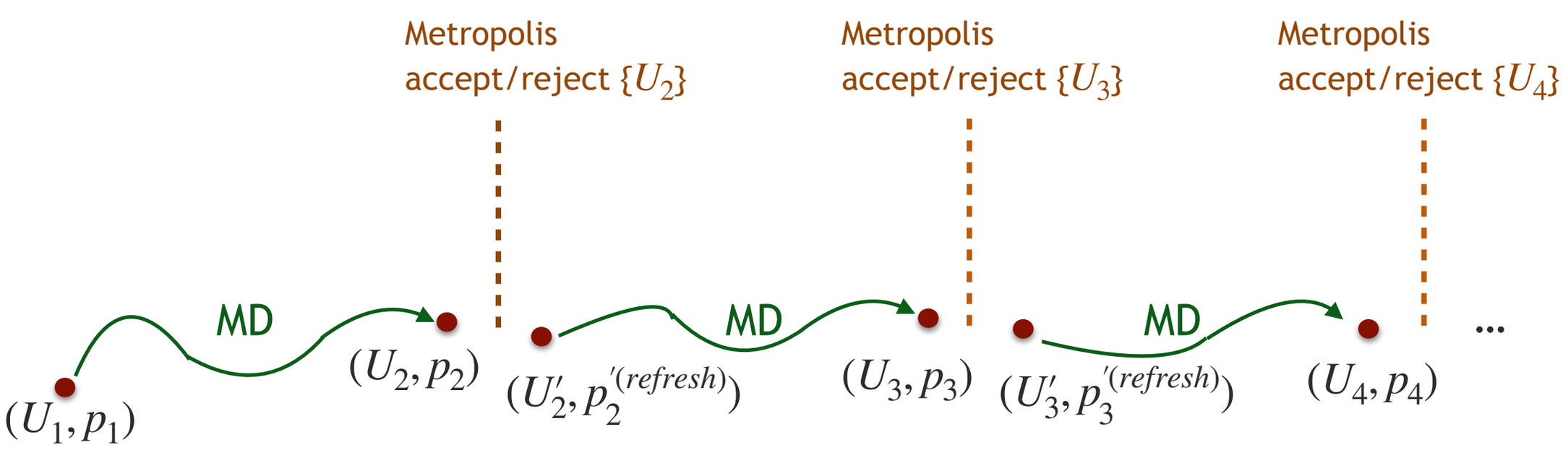
- Use Molecular Dynamics (MD) to generate proposals to Metropolis
- Global acceptance → expect step-size to be driven by number of sites



Metropolis updating via Hybrid Monte Carlo

- Use Molecular Dynamics (MD) to generate proposals to Metropolis
- Global acceptance → expect step-size to be driven by number of sites

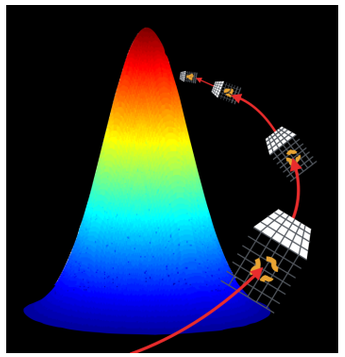
- Multi-time-scale symplectic integrators improve scaling
- Force terms ~ inverse Dirac operator ~ near zero eigenmodes drive physics

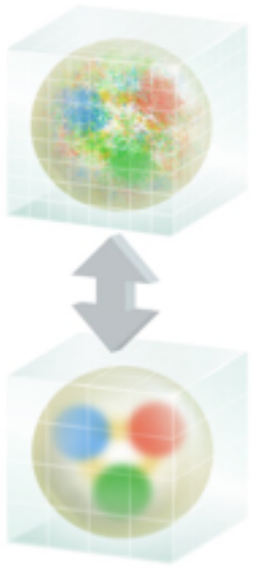


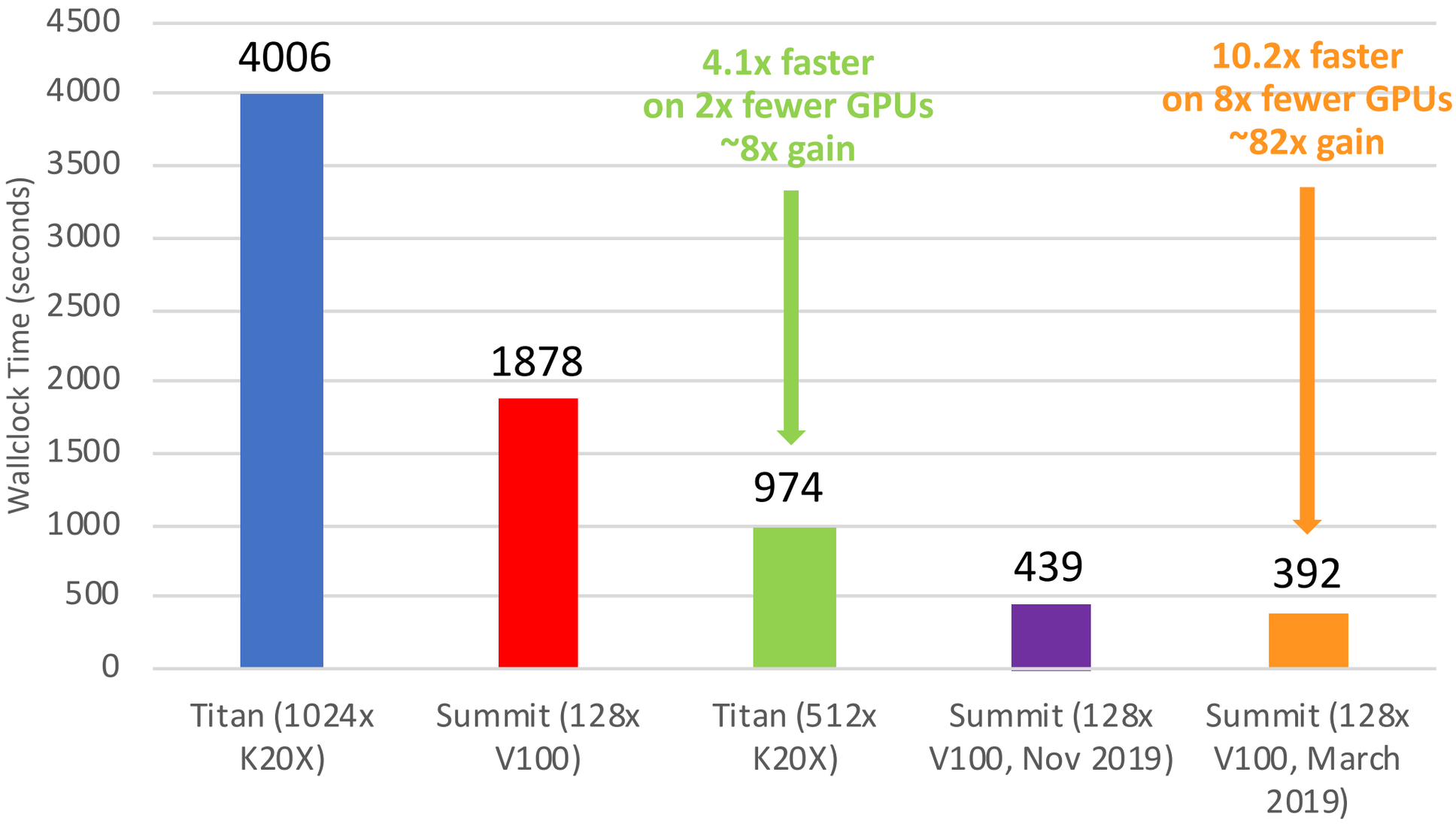
Accelerating QCD gauge generation on GPUs

Collaboration involving ASCR support and Industry partners

Chroma



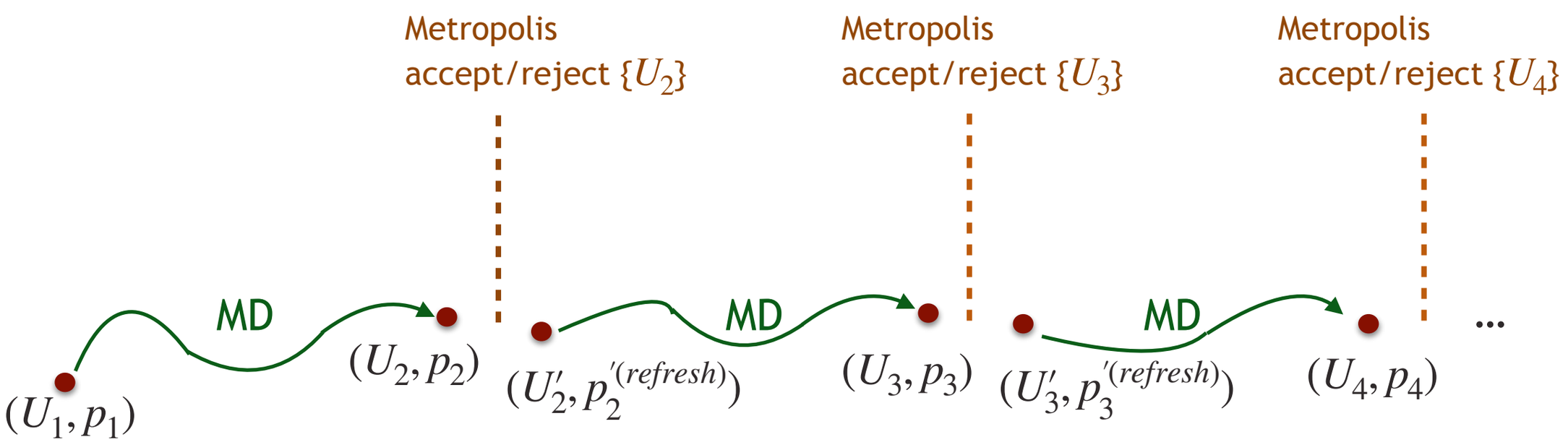




- **~10.2x** wallclock speed-up on Summit using **8x** fewer GPUs than Titan**:**
  ~82x improvement in computational efficiency
- **Allows previously unaffordable calculations**

Partial resolution to critical slowing down

- During integration - solving linear systems
- Multigrid eliminates critical slowing down (ill-conditioning) in the solvers!
- Can use preconditioning techniques - cancellation of force terms
  - Mitigates large condition number in solver



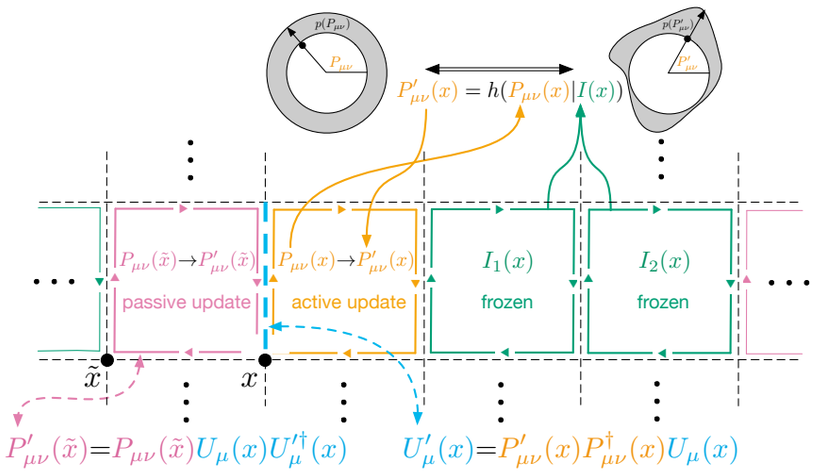
All sounds good so far…

Major goal - eliminate critical slowing down

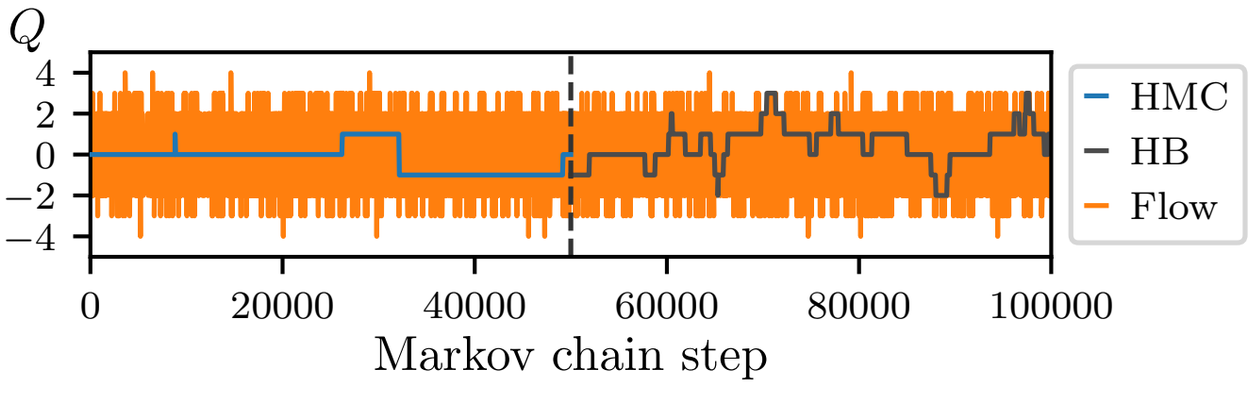
- Small lattice spacing is goal
- Using Molecular Dynamics (MD) to generate proposals to Metropolis
- But not all modes of system evolve at same rate - large autocorrelation time
- Radical solution - use ML methods to “learn” underlying probability distribution
- Does it scale?

2-D toy model with gauge fields

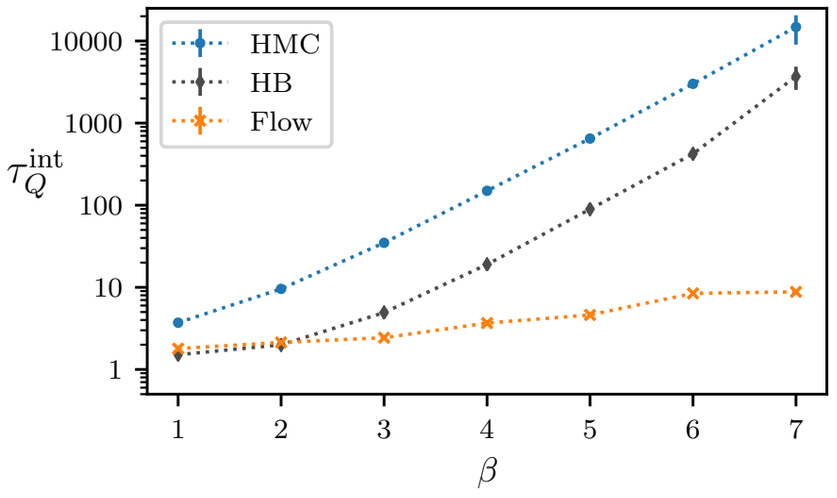
arXiv:2003.06413



Comparison MD, heat-bath, ML flow method



Reduced autocorrelation times



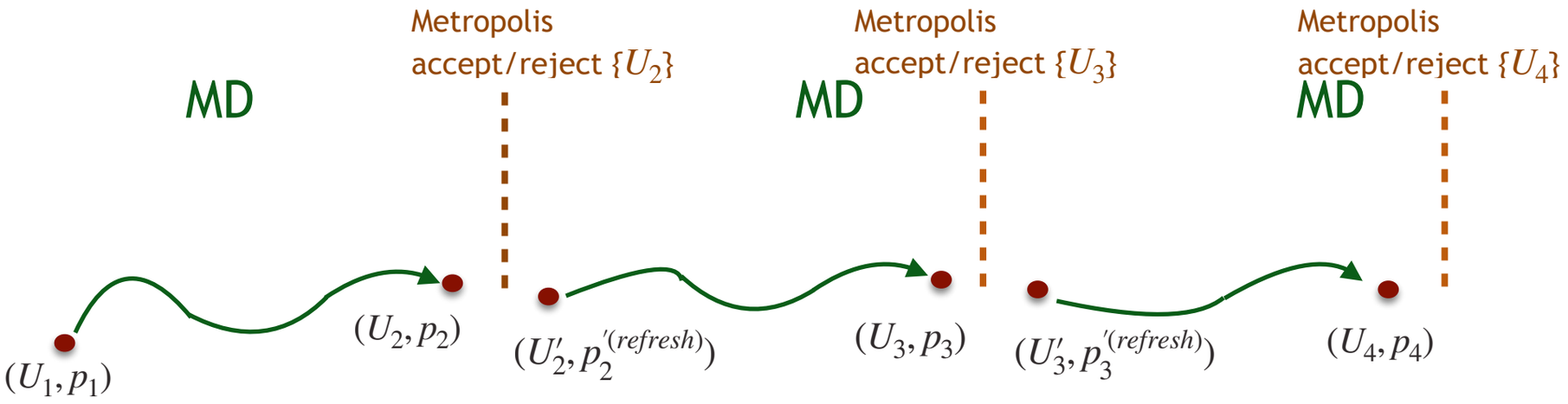
Jefferson Lab

SciDAC-5: tasks - ML

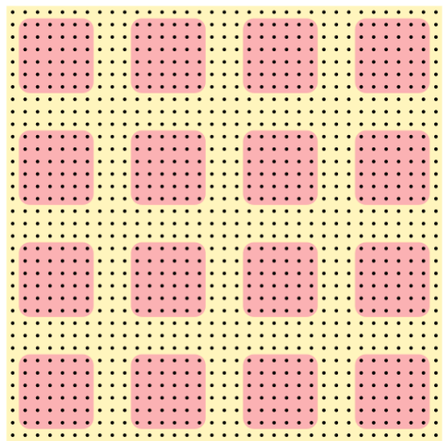
- **ML** (Shanahan)
  - Generative flow (Shanahan)
  - Preconditioning (Balaprakash, Shanahan)
  - Autotuning (Balaprakash, Detmold)
  - Collaborations with ANL, BNL, MIT

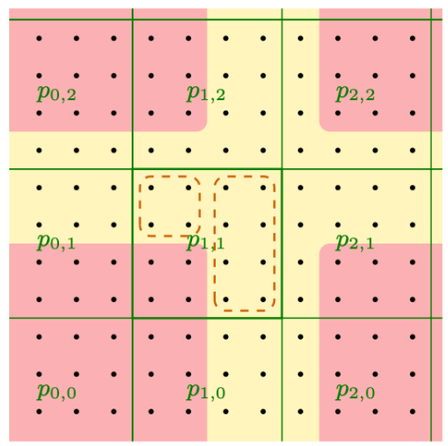
Hierarchial HMC

- Usual HMC updates entire lattice (U fields) in each MD trajectory



- Instead, decompose correlation functions into factors localized to sub-domains





- Integrate subdomains independently
- Increases the effective number of statistical samples

SciDAC-5: tasks - HMC

- **Hierarchial HMC** (Orginos)
  - Need to consider the type of correlation functions & fermion actions
  - Clover (Orginos, Detmold)
  - Staggered and Domain Wall (Mukherjee)
  - Collaborations with JLab, W&M, ORNL, MIT, ANL, BNL

Algorithms for Measurements

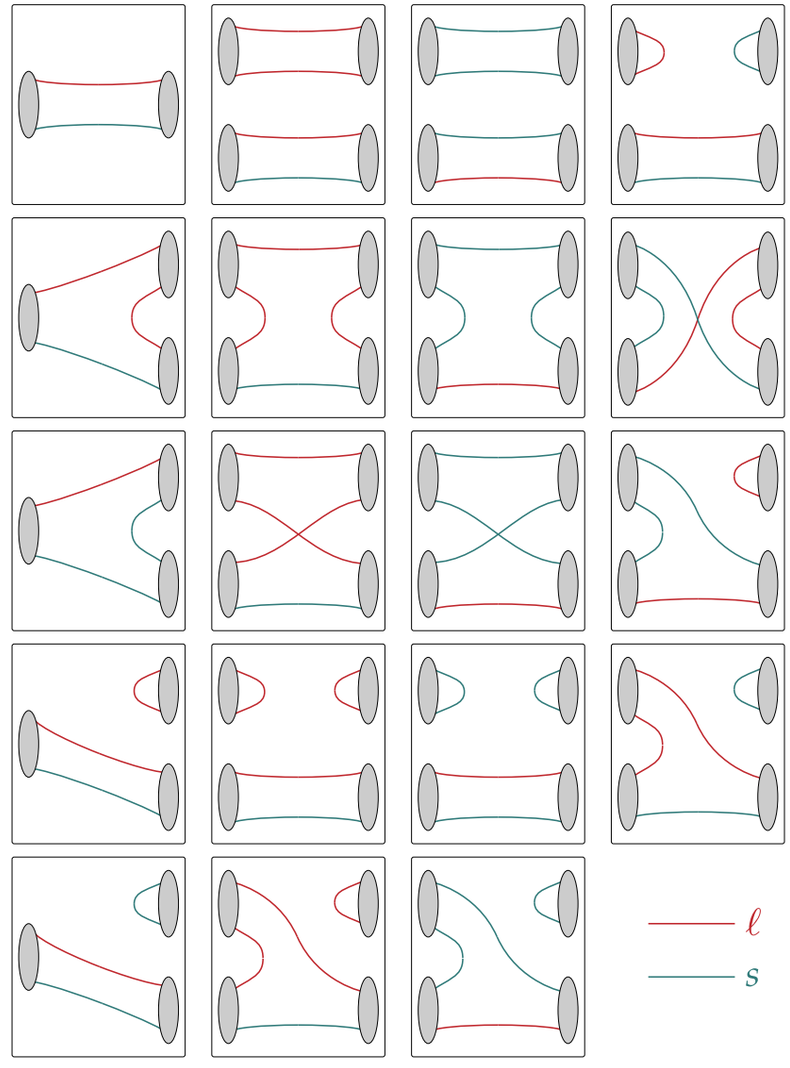
Foundation for LQCD calculations - all on a single gauge field

- Distillation (arXiv:0905.2160)
  - Low-dimensional representation of particles
  - Represent as quark line graphs
  - O(100K- 1M) graphs, O(100K-1M) Dirac solves
  - Supports reuse of propagators

- Generically, requires tensor contractions

- Techniques:
  - Equivalent graph identification
  - Dijkstra's method - optimal walk through space of evaluation order

- Challenge - large ranked objects (think proton, deuteron, etc…)

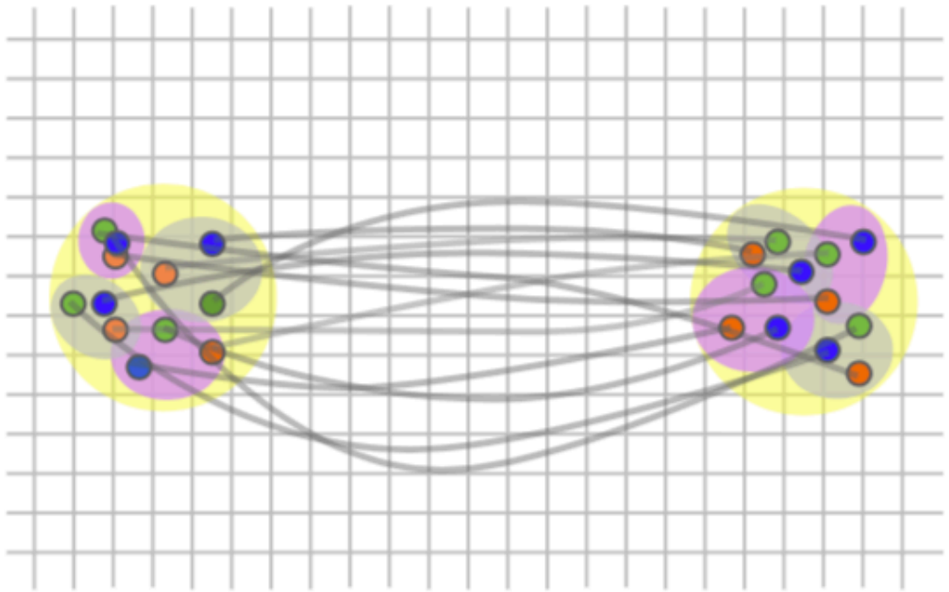


I=1/2 K*π arXiv:1406.4158

QCD for nuclei

- Quarks need to be tied together in all possible way

$N_{contractions} = N_u! N_d! N_s!$ $(\sim 10^{1500}$ for $^{208}Pb)$



- Manage using algorithmic trickery - still significant graph contractions
- TACO @ MIT (c.f. TCE, Cyclops) http://tensor-compiler.org
  - general purpose tensor contraction compiler framework
  - target to QCD specific problems (sparsity patterns, …)
  - Allows for exploration of methodology

Limitations: needs load balancing, flop minimization, data movement…

SciDAC-5: tasks - “Analysis” (Contractions)

- **Analysis** (Detmold)
  - Multinode contractions (Chen)
  - Determinant ordering & blocking (Buluc, Detmold)
  - Sparsity and compression (Amarsinghe, Detmold)
  - Collaborations with JLab, LBNL, MIT

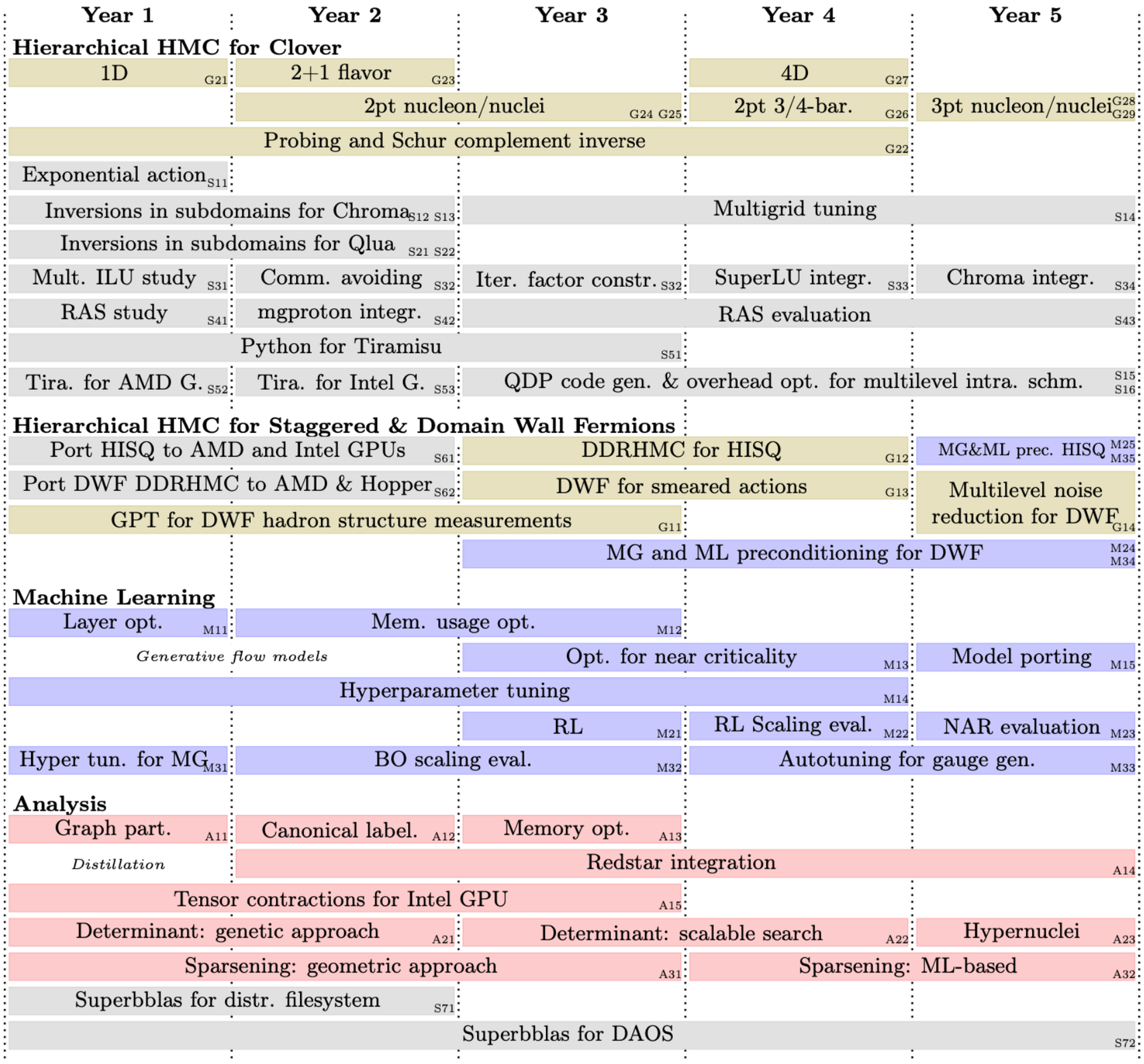
SciDAC-5: tasks - Software

- **Software** (Edwards)
  - Chroma/domains (Winter) - hadron structure
  - Qlua/domains (Pochinsky) - many-body
  - Preconditioning [ILU] (Li)
  - Preconditioning [Schwarz] (Stathopoulos)
  - Tiramisu compiler support (Amarasinghe)
  - Staggered and Domain-Wall porting (Mukherjee)
  - Storage and I/O (Latham, Romero)
  - Collaborations with JLab, WM, LBNL, ORNL, ANL, MIT, BNL, LANL

SciDAC-5: tasks

- **Hierarchial HMC** (Orginos)
  - Clover (Orginos, Detmold)
  - Staggered and Domain Wall (Mukherjee)
  - Collaborations with JLab, W&M, ORNL, MIT, ANL, BNL
- **ML** (Shanahan)
  - Generative flow (Shanahan)
  - Preconditioning (Balaprakash, Shanahan)
  - Autotuning (Balaprakash, Detmold)
  - Collaborations with ANL, BNL, MIT
- **Analysis** (Detmold)
  - Multinode contractions (Chen)
  - Determinant ordering (Buluc, Detmold)
  - Sparsity and compression (Amarasinghe, Detmold)
  - Collaborations with JLab, LBNL, MIT
- **Software** (Edwards)
  - Chroma/domains (Winter) + Qlua/domains (Pochinsky)
  - Preconditioning (Li, Stathopoulos)
  - Tiramisu (Amarasinghe)
  - Staggered and Domain-Wall (Mukherjee)
  - Storage and I/O (Latham, Romero)
  - Collaborations with JLab, WM, LBNL, ORNL, ANL, MIT, BNL, LANL

Timeline


Year 1
Year 2
Year 3
Year 4
Year 5
Hierarchical HMC for Clover
1D G21
2+1 flavor G23
4D G27
2pt nucleon/nuclei G24 G25
2pt 3/4-bar. G26
3pt nucleon/nuclei G28 G29
Probing and Schur complement inverse G22
Exponential action S11
Inversions in subdomains for Chroma S12 S13
Multigrid tuning S14
Inversions in subdomains for Qlua S21 S22
Mult. ILU study S31
Comm. avoiding S32
Iter. factor constr. S32
SuperLU integr. S33
Chroma integr. S34
RAS study S41
mgproton integr. S42
RAS evaluation S43
Python for Tiramisu S51
Tira. for AMD G. S52
Tira. for Intel G. S53
QDP code gen. & overhead opt. for multilevel intra. schm. S15 S16
Hierarchical HMC for Staggered & Domain Wall Fermions
Port HISQ to AMD and Intel GPUs S61
DDRHMC for HISQ G12
MG&ML prec. HISQ M25 M35
Port DWF DDRHMC to AMD & Hopper S62
DWF for smeared actions G13
Multilevel noise reduction for DWF G14
GPT for DWF hadron structure measurements G11
MG and ML preconditioning for DWF M24 M34
Machine Learning
Layer opt. M11
Mem. usage opt. M12
Generative flow models
Opt. for near criticality M13
Model porting M15
Hyperparameter tuning M14
RL M21
RL Scaling eval. M22
NAR evaluation M23
Hyper tun. for MG M31
BO scaling eval. M32
Autotuning for gauge gen. M33
Analysis
Graph part. A11
Canonical label. A12
Memory opt. A13
Distillation
Redstar integration A14
Tensor contractions for Intel GPU A15
Determinant: genetic approach A21
Determinant: scalable search A22
Hypernuclei A23
Sparsening: geometric approach A31
Sparsening: ML-based A32
Superbblas for distr. filesystem S71
Superbblas for DAOS S72

Structure of meeting

- Ordering of talks, and subjects constrained by human schedules/availability
- Goal - speak to milestones in proposal
- Thursday
  - Graphs and contraction techniques
    - Jie Chen, Aydin Buluc, Ryan Abbot, Andrew Pochinsky
  - High performance storage
    - Robert Latham
  - Group dinner this evening - aiming for 6pm @ “Aago” (Warrick Blvd)

- Friday
  - ML
    - Phiala Shanahan, Prasanna Balaprakash
  - HMC
    - Dennis Bollweg, Kostas Orginos, Eloy Romero
  - Software
    - Andreas Stathopoulos, Sherry Li, Frank Winter, Teo Collin, Balint Joo
  - Discussion/organization of efforts