# Large Domain DDHMC

Peter Boyle (BNL),
Christopher Kelly (BNL),
Azusa Yamaguchi (University of Edinburgh)
RBC-UKQCD collaboration

- Large domain DDHMC for GPUs
  - Goal: enable 2+1+1f simulations, 4GeV cut off

# The RBC & UKQCD collaborations

*UC Berkeley/LBNL*
Aaron Meyer

*BNL and BNL/RBRC*
Yasumichi Aoki (KEK)
Peter Boyle (Edinburgh)
Taku Izubuchi
Yong-Chull Jang
Chulwoo Jung
Christopher Kelly
Meifeng Lin
Hiroshi Ohki
Shigemi Ohta (KEK)
Amarjit Soni

*CERN*
Andreas Jüttner (Southampton)

*Columbia University*
Norman Christ
Duo Guo
Yikai Huo
Yong-Chull Jang
Joseph Karpie
Bob Mawhinney
Ahmed Sheta
Bigeng Wang
Tianle Wang
Yidi Zhao

*University of Connecticut*
Tom Blum
Luchang Jin (RBRC)
Michael Riberdy
Masaaki Tomii

*Edinburgh University*
Matteo Di Carlo
Luigi Del Debbio
Felix Erben
Vera Gülpers
Tim Harris
Raoul Hodgson
Nelson Lachini
Michael Marshall
Fionn Ó hÓgáin
Antonin Portelli
James Richings
Azusa Yamaguchi
Andrew Z.N. Yong

*KEK*
Julien Frison

*University of Liverpool*
Nicolas Garron

*Michigan State University*
Dan Hoying

*Milano Bicocca*
Mattia Bruno

*Peking University*
Xu Feng

*University of Regensburg*
Davide Giusti
Christoph Lehner (BNL)

*University of Siegen*
Matthew Black
Oliver Witzel

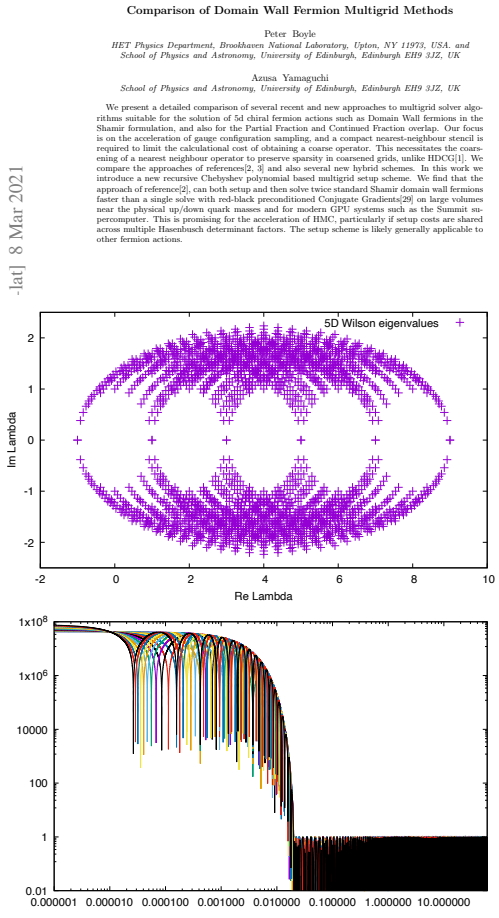
*University of Southampton*
Nils Asmussen
Alessandro Barone
Jonathan Flynn
Ryan Hill
Rajnandini Mukherjee
Chris Sachrajda

*University of Southern Denmark*
Tobias Tsang

*Stony Brook University*
Jun-Sik Yoo
Sergey Syritsyn (RBRC)

# Domain Wall Multigrid

- Preprint: https://arxiv.org/pdf/2103.05034.pdf
- Spectrum of DWF makes coarsening nearest neighbour operator *hard*
  - Polynomial approximation to $\frac{1}{z}$ in region of complex plane enclosing origin
  - Typically solve normal equations on positive definite $M^\dagger M$
  - Nearest neighbour coarsenings of $\gamma_5 R_5 D_{dwf}$ (Herm, indefinite)
  - Coarse pace preserves $\Gamma_5 = \gamma_5 R_5$: use $1 \pm \Gamma_5$ projected subspace vectors
- Novel chebyshev polynomial setup of multigrid
- Result:
  **Set up and solve twice $D_{dwf}$ faster than red-black CG**
- HMC focus; use compressed Lanczos for valence analysis




Beware false baseline papers: using unpreconditioned CG is a bad baseline, especially for Wilson.

Beware papers counting fine matrix multiples without time for coarse space.

# QCD path integral

- Partition function becomes a real, statistical mechanical probability weight

$$Z = \int d\bar{\psi} d\psi dU e^{-S_G[U] - S_F[\bar{\psi}, \psi, U]}$$

- Dirac differential operator represented via discrete derivative approximations: sparse matrix
- Use pseudofermion approach to replace with Gaussian integral $\sqrt{\pi\lambda} = \int dt e^{-t^2/\lambda}$

$$\int \mathcal{D}\bar{\psi}\mathcal{D}\psi e^{-\bar{\psi}(x) A_{xy} \psi(y)} = \det A$$

$$\pi\lambda = \int d\phi_r e^{-\phi_r \frac{1}{\lambda} \phi_r} \int d\phi_i e^{-\phi_i \frac{1}{\lambda} \phi_i} = \int d\phi^* d\phi e^{-\phi^* \frac{1}{\lambda} \phi}$$

- replace two flavour determinant with a two flavour *pseudofermion* integral

$$(\det M)^2 = (\det \gamma_5 M)^2 = \det M^\dagger M = \int \mathcal{D}\phi^* \mathcal{D}\phi e^{-\phi^*(x)(M^\dagger M)^{-1}\phi(y)}$$

# Hybrid Monte Carlo

- Auxiliary Gaussian integral over conjugate momentum field $\int d\pi e^{\frac{-\pi^2}{2}}$
  Lives in Lie algbra; serves only to move $U$ round the group Manifold

$$\int d\pi \int d\phi \int dU \quad e^{-\frac{\pi^2}{2}} e^{-S_G[U]} e^{-\phi^*(M^\dagger M)^{-1}\phi}$$

- Outer Metropolis Monte Carlo algorithm
  - Draw momenta
  - Draw pseudofermion as gaussian $\eta = M^{-1}\phi$
  - Metropolis acceptance step
- Metropolis proposal includes inner molecular dynamics at constant Hamiltonian:

$$H = \frac{\pi^2}{2} + S_G[U] + \phi^*(M^\dagger M)^{-1}\phi$$

$$\dot{U} = i\pi U \qquad ; \qquad i\dot{\pi} = -(U\nabla_U S)_{TA}$$

- Must invert $M^\dagger M$ at each timestep of evolution in MD force

$$\delta(M^\dagger M)^{-1} = -(M^\dagger M)^{-1}[(\delta M^\dagger)M + M(\delta M)](M^\dagger M)^{-1}$$

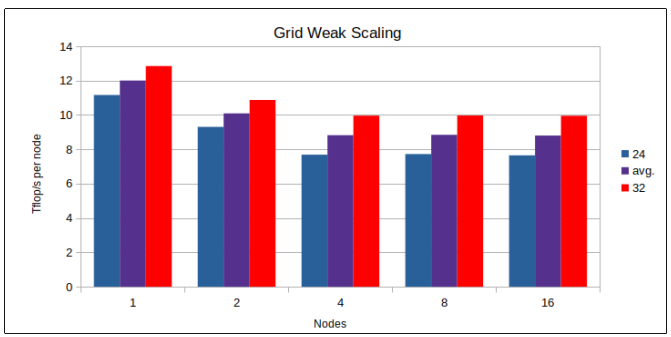
# Large domain DDHMC for GPUs

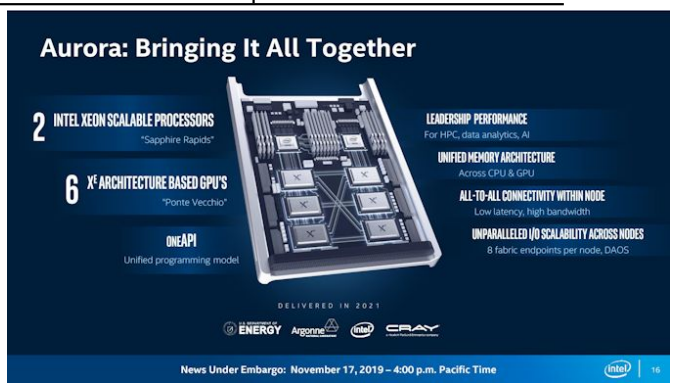
**Motivation:**

- GPU speed is increasing rapidly over time
- Interconnect speeds are *not* keeping pace.
- Expense spent on interconnect is significant
- If we get nearer commodity pricing, this will be even worse
- Project 30% efficiency on future systems
- GPU cache sizes are growing
  - DWF reuses gauge links $L_s$ times
  - DWF reuses spinors $2N_d$ times
  - $\Rightarrow$ cache bound performance on a single node
- **Aurora:**
  - over 130TF/s fp64 (https://www.alcf.anl.gov/aurora)
  - "Rambo cache"; Xe memory fabric (disclosed by Intel, HotChips conference ) (https://www.nextplatform.com/2020/09/02/intel-puts-its-xe-gpu-stakes-in-the-ground/)
- Interconnect will rapidly become bottleneck

# System balance

| System | GPUs | Node peak FP32 | Node interconnect (GB/s Snd+Rcv) |
|---|---|---|---|
| Booster/Jülich | 4 x A100 | 78TF/s | 200GB/s |
| Tursa/Edinburgh | 4 x A100 | 78TF/s | 200GB/s |
| Summit | 6 x V100 | 94TF/s | 50GB/s |
| Aurora | 6 x Intel Xe | $\geq$ 130 TF/s fp64<br>$\geq$ 260 TF/s fp32 (?)<br>(conjectured 2x) | 300-400GB/s<br>300GB/s to GPU's? |




Future machines will not scale with current algorithms

# System balance

32^4 comms and
Compute perfectly
Overlapped

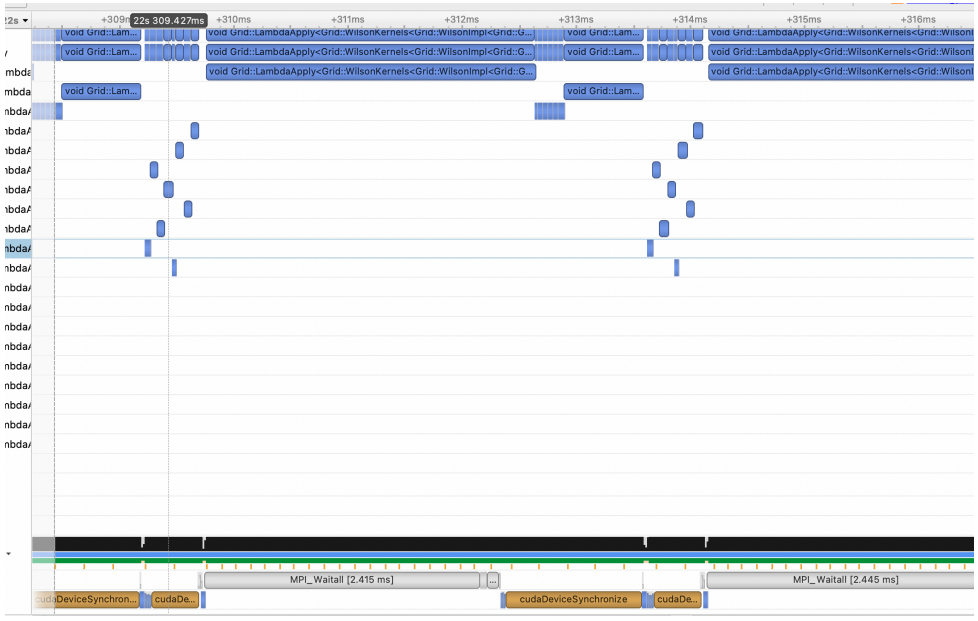
Stopped using Nvlink
With GPU, use RDMA

Read coalesce kernels

750us -> 60us

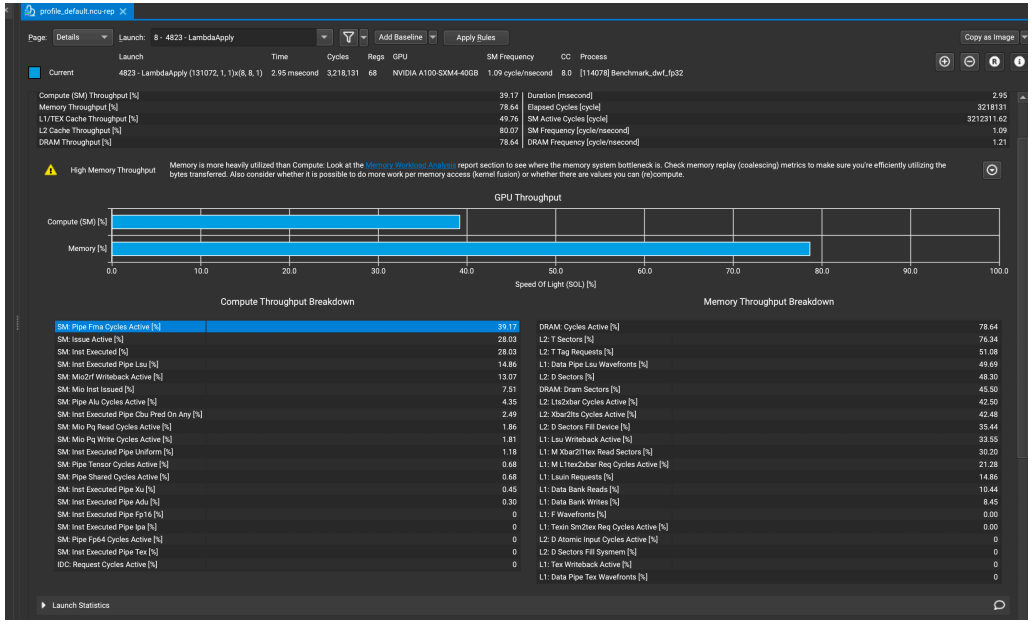
6.6TF/s -> 9.9 TF/s

Grid benchmark
5.3 last Nov, now 9.9
With many improvements



# System balance

Dslash kernel: 39% FMA pipe, 80% L2, 78% memory; hard to improve by much



# Subdomains

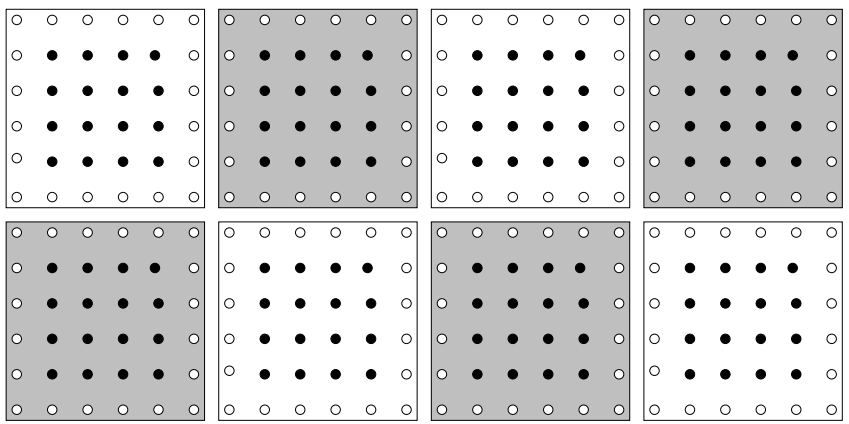
We decompose space time into hypercuboidal blocks of size $L^4$.
The block coordinate is (in integer division): $b_i = x_i / L$
The intra block coordinate is: $l_i = x_i | L$.
We assign to each block a parity: $p = (\sum_i b_i) | 2$.
Define domains $\Omega$ and $\bar{\Omega}$ as the set of points within blocks of parity zero and parity one respectively.
*exterior* boundary haloes are $\partial_\Omega$ and $\partial_{\bar{\Omega}}$ such that,

$$\partial_\Omega \cap \Omega = \emptyset,$$

and

$$\partial_{\bar{\Omega}} \cap \bar{\Omega} = \emptyset,$$

respectively.



The Dirac operator may then be written as

$$D = \begin{pmatrix} D_\Omega & D_\partial \\ D_{\bar{\partial}} & D_{\bar{\Omega}} \end{pmatrix}.$$

# DDHMC refresher

Fermion operator may be factored:

$$\begin{pmatrix} D_\Omega & D_\partial \\ D_{\bar\partial} & D_{\bar\Omega} \end{pmatrix} = \begin{pmatrix} 1 & D_\partial D_{\bar\Omega}^{-1} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} D_\Omega - D_\partial D_{\bar\Omega}^{-1} D_{\bar\partial} & 0 \\ 0 & D_{\bar\Omega} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ D_{\bar\Omega}^{-1} D_{\bar\partial} & 1 \end{pmatrix}. \quad (1)$$

The factors $L$, $M$, and $U$ are obvious and the determinant is:

$$\det D = \det D_\Omega \det D_{\bar\Omega} \det \left\{ 1 - D_\Omega^{-1} D_\partial D_{\bar\Omega}^{-1} D_{\bar\partial} \right\},$$

# DDHMC refresher

- hep-lat/0409106
- Partition the lattice into hypercuboids
- Colour them black and white according to parity
- Call "white" domain $\Omega$ and complement $\bar{\Omega}$
- Schur factoring the Fermion determinant leaves local and non-local terms that can be integrated on different timescales.

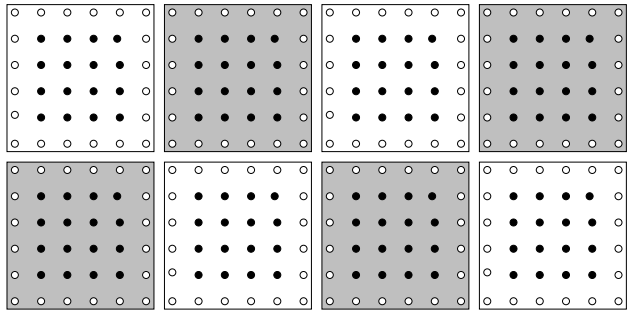
$$D = \begin{pmatrix} D_\Omega & D_\partial \\ D_{\bar{\partial}} & D_{\bar{\Omega}} \end{pmatrix}$$

$$\det D = \det D_\Omega \det D_{\bar{\Omega}} \det \left\{ 1 - D_\Omega^{-1} D_\partial D_{\bar{\Omega}}^{-1} D_{\bar{\partial}} \right\},$$

Schwarz-preconditioned HMC algorithm
for two-flavour lattice QCD

Martin Lüscher

*CERN, Physics Department, TH Division*
*CH-1211 Geneva 23, Switzerland*



- *small* domains $4^4$ to $6^4$
- HMC MD integrate gauge action and local determinants for each domain without communication
- Fits within L2 cache of a CPU core
- Small cell provides IR regulator for Dirichlet Dirac solves
- Exterior boundary gauge links are frozen (cross domain and in surface plane)

# Boundary determinant

- Handling the Schur complement "boundary" determinant requires care

$$\chi = 1 - D_\Omega^{-1} D_\partial D_{\bar\Omega}^{-1} D_{\bar\partial}$$

- Can restrict to exterior boundary of $\Omega$

$$R = \mathbb{P}_{\bar\partial} - \mathbb{P}_{\bar\partial} D_\Omega^{-1} D_\partial D_{\bar\Omega}^{-1} D_{\bar\partial}$$

- because in the right basis $\chi$ takes the form

$$\chi = \begin{pmatrix} 1 - X & 0 \\ Y & 1 \end{pmatrix}$$

  so $\det \chi = \det R = \det(1 - X)$

- For pseudofermion action $\phi_{\bar\partial}^\dagger (RR^\dagger)^{-1} \phi_{\bar\partial}$,

$$R^{-1} = \hat{\mathbb{P}}_{\bar\partial} - \hat{\mathbb{P}}_{\bar\partial} D^{-1} \hat{D}_{\bar\partial}$$

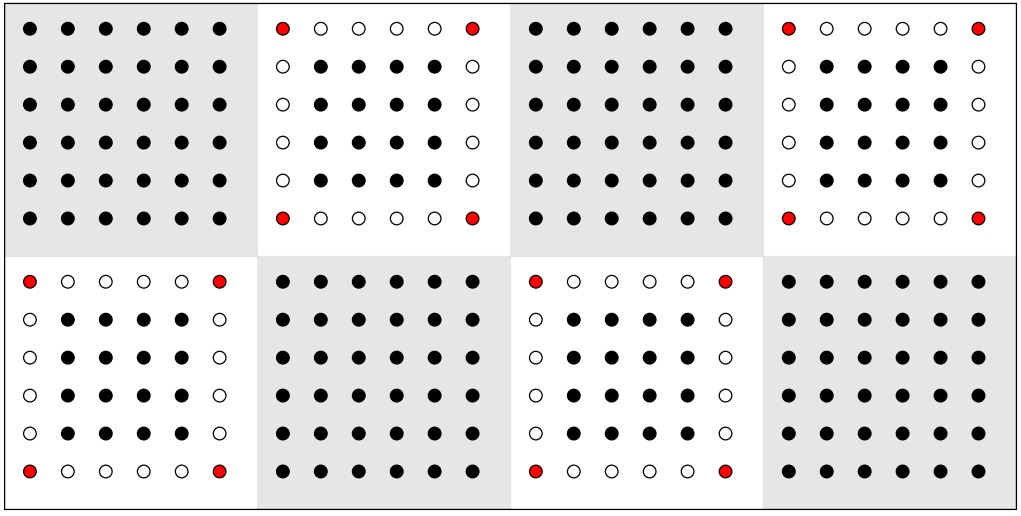
- $\delta R^{-1} = \mathbb{P}_{\bar\partial} D^{-1} \delta D D^{-1} D_{\bar\partial}$.
- Pauli-Villars (or Hasenbusch) requires

$$\phi_{\bar\partial}^\dagger P^\dagger R^{-\dagger} R^{-1} P \phi_{\bar\partial}.$$

  and $\delta R = \mathbb{P}_{\bar\partial} D_\Omega^{-1} (\delta D_\Omega) D_\Omega^{-1} D_\partial D_{\bar\Omega}^{-1} D_{\bar\partial} + \mathbb{P}_{\bar\partial} D_\Omega^{-1} D_\partial D_{\bar\Omega}^{-1} (\delta D_{\bar\Omega}) D_{\bar\Omega}^{-1} D_{\bar\partial}$.

# Symmetric domain shapes

- Luscher's domain structure



- Boundary pseudofermion lives on the interior boundary of $\Omega$
- Spin structured: sites on only one face are spin projected
  - Red dots are four component pseudofermion
  - Open dots are two component pseudofermion

# Large domain DDHMC

- GPU's offer large parallelism within the node
  - $32^4$ or greater subvolume per domain
  - Local solves can outstrip the network.
  - Node 10x(?) faster than Booster, network 1.5x faster (?)
  - Domain decompose HMC on *large* domains
- Cell local Dirichlet determinants are "obvious"
  - Created an adaptor for any Grid Fermion operator that zeroes gauge links, removes communication
  - Standard two flavour pseudofermion action otherwise.
  - Local determinant equally ill conditioned as light solve
    $\longrightarrow$ this is exactly what GPU's are good at!
- *Will also change subdomain shapes*
- Want maximal domain size on each node, and load balanced

# Domain Wall force

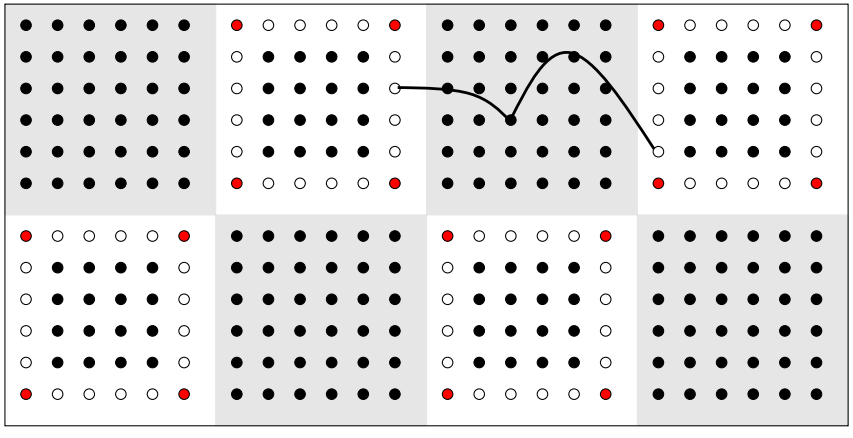
- Normal equations on 5D system uses single solve in force

$$\phi^\dagger (M^\dagger M)^{-1} \phi$$

  - Can also be used for local determinant
- Boundary projector means number of solves is doubled (normal equations twice)

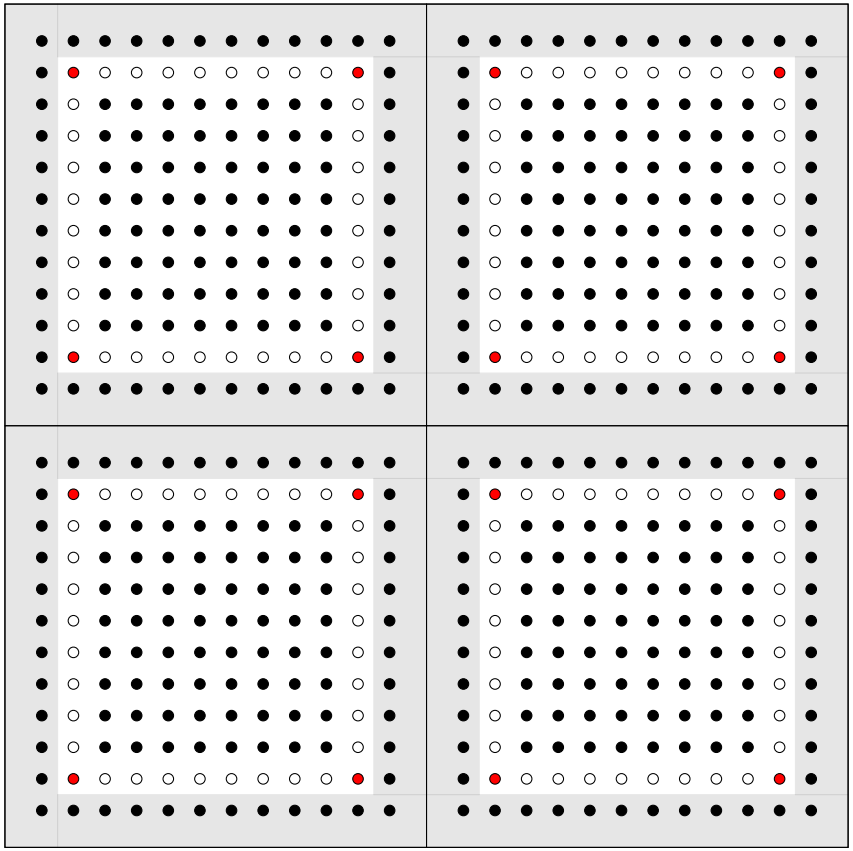
$$\delta \left( \phi_{\bar{\partial}}^\dagger R^{-\dagger} R^{-1} \phi_{\bar{\partial}} \right) = 2\mathrm{Re} \langle R^{-1} \phi_{\bar{\partial}} | \mathbb{P}_{\bar{\partial}} D^{-1} \delta D D^{-1} D_{\bar{\partial}} \phi_{\bar{\partial}} \rangle$$

- Must have a good integrator timestep ratio between local and boundary determinants
- Force is suppressed by two light quark propogators
  - Can suppress force arbitrarily by using a broader band of inactive links
  - Short distance propagator is not dictated by pion mass



# Non-symmetric domain shapes

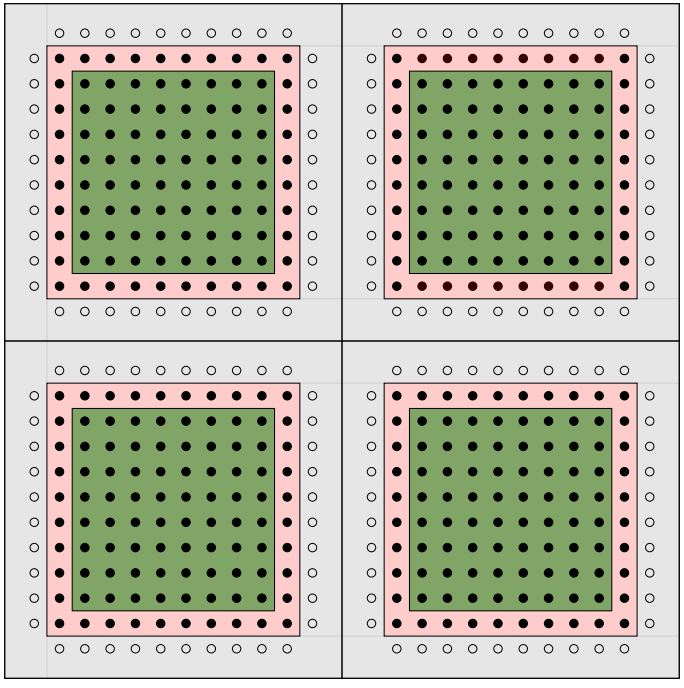
$$S_{\mathrm{Pseudofermion}} = \phi^{\dagger}_{\Omega}(D^{\dagger}_{\Omega}D_{\Omega})^{-1}\phi_{\Omega} + \phi^{\dagger}_{\bar{\partial}}(R^{\dagger}R)^{-1}\phi_{\bar{\partial}}$$

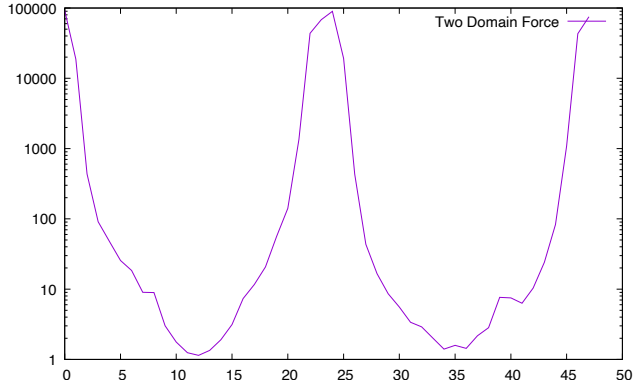


Large Domain DDHMC domain structure

- Boundary pseudofermion lives on the interior boundary of $\Omega$ (or $\bar{\Omega}$)
- $\mathrm{det}D_{\Omega}$ is local to a node and maximally large
- Freeze all links in $\bar{\Omega}$, *do not need to compute* $\mathrm{detD}_{\bar{\Omega}}$

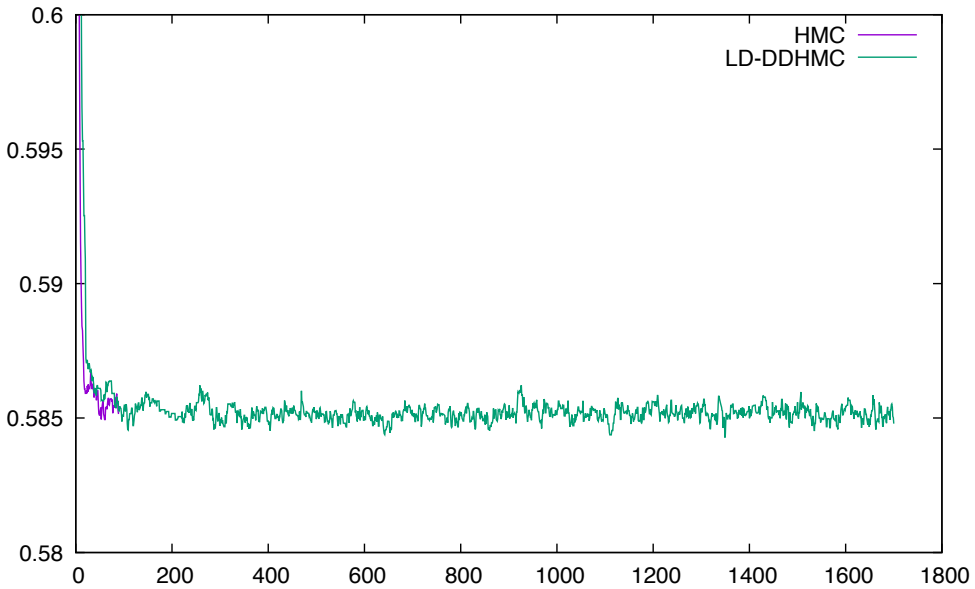

- HMC slow zone close to boundary
  - scale the HMC time evolution in a coordinate dependent way
  - power law "slow down" in red zone
  - Counterbalance rise in propagator





# $16^3 \times 48$ 2f test

- DWF+Iwasaki 2 flavor: $\beta = 2.13, 16^3 \times 48, m_f = 0.01, L_s = 16$
- Produced on 2 GPUs
- 3:1 ratio of boundary determinant to local determinant timesteps
  - Omelyan integrator (2 force evaluations in nesting)
  - Adequate hierarchy in integration
- Wall clock gain depends on interconnect performance: substantial factor on Aurora, Perlmutter, Summit
- Strange quark / odd flavours are a work in progress. May just use EOFA.




Clearly need more statistics on reference, but looks OK
Evolution is solid and plaquette in low stats agreement with

- 2f Grid run of same ensemble
- Plausibly close to historic 2+1f u/d/s plaquette

# Summary and outlook

- Large domain DDHMC is ideal to decouple islands of high performance in future GPU systems
  - Conjecture up to 8x acceleration of local domain solves (?)
  - Precise, algorithmically efficient determinant factorisation:
- Can also consider multilevel integration
  - no in-principle barrier for DWF
  - $N^2$ valence measurements
    $\Rightarrow$ For DWF need a better valence solver scheme as we have not yet achieved the Wilson multigrid speed-up
    $N^2$ Lanczos deflation prohibitive
- Fall of propagator with distance and computer architecture trends make this a guaranteed win in long run
  Expect it to win on Aurora