Disconnected contributions from GPU's

Alejandro Vaquero

Computational-based Science and Technology Research Center (CaSToRC) at The Cyprus Institute

In collaboration with:

Abdou Abdel-Rehim, CaSToRC at The Cyprus Institute
Constantia Alexandrou, CaSToRC and University of Cyprus
Giannis Koutsou, CaSToRC at The Cyprus Institute
Alexei Strelchenko, CaSToRC at The Cyprus Institute

Martha Constantinou, University of Cyprus
Kyriacos Hadjiyiannakou, University of Cyprus

Simon Dinter, DESY, NIC
Vincent Drach, DESY, NIC
Karl Jansen, DESY, NIC

October 16th, 2012

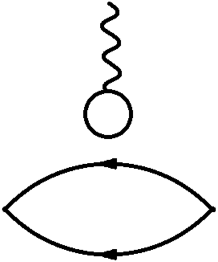
THE CYPRUS INSTITUTE

CaSToRC

Outline

- Brief introduction to disconnected contributions
- Stochastic procedures
- Truncated Solver Method (TSM)
- The one-end trick and other improvements
- GPU performance and scaling
- The summation method
- Results
- Conclusions and future plans

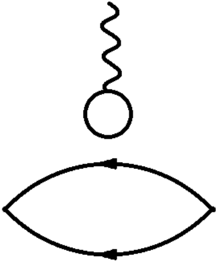
Motivation



- **Determination of flavour singlet quantities**
  η mass, nucleon form factors...
- **Computations of non-perturbative nature**
- **We must rely on lattice methods**

$$L(x) = \mathrm{Tr}\left[\Gamma G(x;x)\right]$$

Disconnected contributions



$L(x) = \mathrm{Tr}\left[\Gamma G(x;x)\right]$

- For the evaluation of disconnected diagrams we need to compute all-to-all propagators
- Very expensive from the computational point of view
- Neglected in most hadron structure studies

Stochastic procedures

- **Exact computation of the all-to-all unfeasible nowadays**
- **We can use stochastic techniques**
  - Invert a random set of sources $|\eta_j\rangle$ that form a basis up to stochastic errors
  - Properties $\begin{cases} \frac{1}{N}\sum_{j=1}^{N} |\eta_j\rangle = O\left(\frac{1}{\sqrt{N}}\right) \\ \frac{1}{N}\sum_{j=1}^{N} |\eta_j\rangle\langle\eta_j| = I + O\left(\frac{1}{\sqrt{N}}\right) \end{cases}$
  - In this work we use $\mathbf{Z}_2$ and $\mathbf{Z}_4$ noise sources
- **So we get an unbiased estimation of the all-to-all propagator**

$$M|s_j\rangle = |\eta_j\rangle \longrightarrow M_E^{-1} := \frac{1}{N}\sum_{j=1}^{N} |s_j\rangle\langle\eta_j| \approx M^{-1}$$

Stochastic procedures

- **Error decresases as $1/\sqrt{N}$, we usually need a large number of stochastic sources N**
- **Each stochastic source requires an inversion of the fermionic matrix**

$$M\,|s_j\rangle = |\eta_j\rangle$$

- **To reduce the stochastic noise, we want to increase N at a reduced cost**
- **The Truncated Solver Method (TSM) allows us to do this**

The Truncated Solver Method

- **First published in PoSLaT2007, 141 (G. Bali, S. Collins and A. Schäffer)**
- **Instead of solving $M\,|s_j\rangle = |\eta_j\rangle$ exactly, we aim at a low precision estimation**
  - Cut the inverter (CG) at a certain number of iterations OR at a given precision $\rho^2 \sim 10^{-4}$
- **This is cheap (fast inversions) but inaccurate $\longrightarrow$ We introduce a bias**
- **We compute the correction of the bias stochastically**

$$M_E^{-1} := \frac{1}{N_{HP}} \sum_{j=1}^{N_{HP}} \left(|s_j\rangle \langle \eta_j|_{HP} - |s_j\rangle \langle \eta_j|_{LP} \right) + \frac{1}{N_{LP}} \sum_{j=N_{HP}+1}^{N_{HP}+N_{LP}} |s_j\rangle \langle \eta_j|_{LP}$$

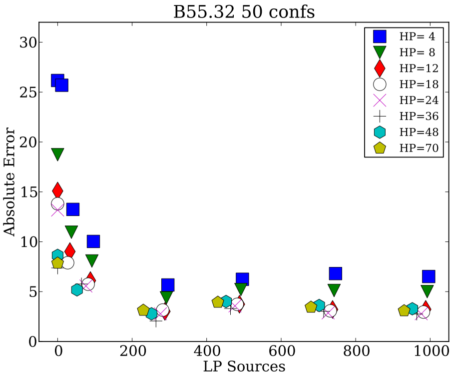
The Truncated Solver Method

- **If the convergence in the inversions is fast, we can get away with a low N_{HP}**
- **Error should decrease essentially as $1/\sqrt{N_{LP}}$**
- **Since the LP sources don't require an accurate inversion, we can take advantage of the half precision algorithms for GPU's**

  Mixed double/single ~ 100 GFlops **Mixed double/half ~ 170 GFlops**

- **Two parameters to tune: precision of LP and N_{HP}/N_{LP} ratio**
- **Fine-tuning depends on the loop to be computed**

Determination of the TSM parameters



- Data for $\bar{\psi}\gamma_3 D_3 \psi$, a piece of $\langle x \rangle_{u+d}$
- After 300 LP, hard to improve
- 24HP/300LP$\approx$ 48HP are enough for all loops with local and one-derivative insertion

The one-end trick

- For twisted mass fermions, the difference of propagators in the twisted basis is

$$M_u^{-1} - M_d^{-1} = -2i\mu M_d^{-1}\gamma_5 M_u^{-1}$$

- So, instead of computing the l.h.s. we do the r.h.s. as

$$\sum X\left(M_u^{-1} - M_d^{-1}\right) = -2i\mu \sum_r \left\langle s^\dagger X \gamma_5 s \right\rangle_r$$

- In principle, the trick only works for the difference, but an alternative version can be developed for the sum

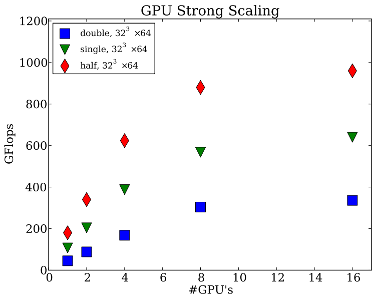
$$\sum X\left(M_u^{-1} + M_d^{-1}\right) = 2\sum_r \left\langle s^\dagger \gamma_5 X \gamma_5 D_W s \right\rangle_r$$

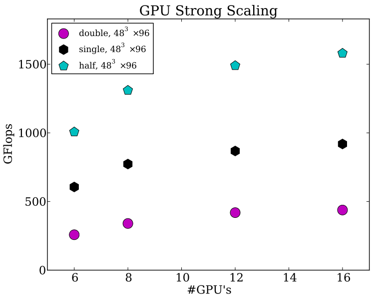
The one-end trick

- **Below a list of bilinears with its appropiate version of the one-end trick**

Bilinear	Twisted Basis	Standard vv	Generalized vv
$\bar{\psi}\psi$	$i\bar{\psi}\gamma_5\tau_3\psi$	✓	✗
$\bar{\psi}\tau_3\psi$	$i\bar{\psi}\gamma_5\psi$	✗	✓
$i\bar{\psi}\gamma_5\psi$	$-\bar{\psi}\tau_3\psi$	✓	✗
$i\bar{\psi}\gamma_5\tau_3\psi$	$-\bar{\psi}\psi$	✗	✓
$\bar{\psi}\gamma_\mu\psi$	$\bar{\psi}\gamma_\mu\psi$	✗	✓
$\bar{\psi}\gamma_5\gamma_\mu\psi$	$\bar{\psi}\gamma_5\gamma_\mu\psi$	✗	✓
$\bar{\psi}\gamma_\mu D_\nu\psi$	$\bar{\psi}\gamma_\mu D_\nu\psi$	✗	✓
$\bar{\psi}\gamma_\mu D_\nu\tau_3\psi$	$\bar{\psi}\gamma_\mu D_\nu\tau_3\psi$	✓	✗
$\bar{\psi}\gamma_5\gamma_\mu D_\nu\psi$	$\bar{\psi}\gamma_5\gamma_\mu D_\nu\psi$	✗	✓
$\bar{\psi}\gamma_5\gamma_\mu D_\nu\tau_3\psi$	$\bar{\psi}\gamma_5\gamma_\mu D_\nu\tau_3\psi$	✓	✗

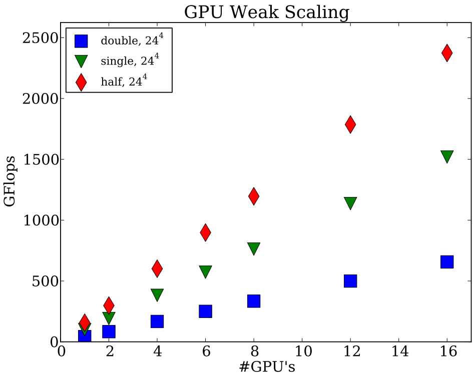
GPU performance and scaling





- Strong scaling competitive up to 8 gpu
- Strongly depends on local volume

GPU performance and scaling



- **Almost perfect weak scaling up to 16 gpu**

Other improvements

- **Generation of noise sources on-the-fly**
  - Don't store propagators/sources, saving I/O - storage (very important for LP sources)
- **Implementation of contractions directly on GPU's**
  - We take advantage of a massively parallel architecture
- **Usage of cudaFFT library to generate all momenta**

The summation method

- **Alternative to the plateau method for computing ratios**
- **Excited states suppressed by e^{-mt_s}, instead of e^{-mt_i}**
- **Requires the 3pt at several $t_s \rightarrow$ more expensive for the connected**
  - L. Maiani, G. Martinelli, M. L. Paciello, B. Taglienti, Nucl. Phys. B**293**, 420 (1987)
  - S. Güsken, arXiv:hep-lat/9906034v1
  - S. Capitani, B. Knippschild, M. Della Morte, H. Wittig, PoSLaT2010 147
- **The 3pt is summed over t_i up to t_{Sink}**

$$R_{SUM}(t_s) = \sum_{t_i=0}^{t_s} R_{PLA}(t_s, t_i)$$

The summation method

- **We usually require $t_i >> 1$ and $t_s - t_i >> 1$ in order to remove undesirable contributions from excited states**

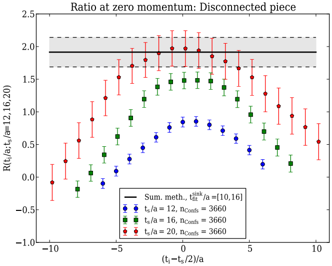
$$R_{PLA}(t_s, t_i) = R_{GS} + O\left(e^{-Kt_i}\right) + O\left(e^{-K'(t_s - t_i)}\right)$$

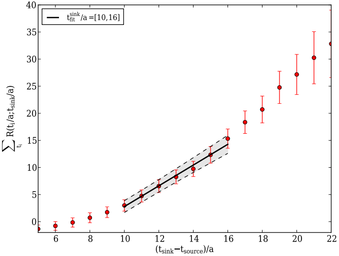
- **However, if we sum up to t_s, the unwanted contributions form a geometric series and become**

$$R_{SUM}(t_s) = t_s R_{GS} + c\left(K, K'\right) + O\left(e^{-Kt_s}\right) + O\left(e^{-K't_s}\right)$$

- **This way the contribution of excited states is always supressed by an e^{-Kt_s} factor**

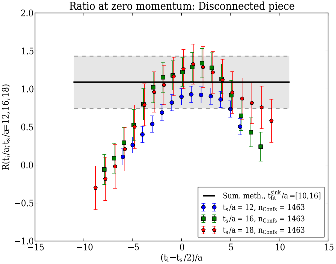
Results: $\sigma_{\pi N}$ disconnected

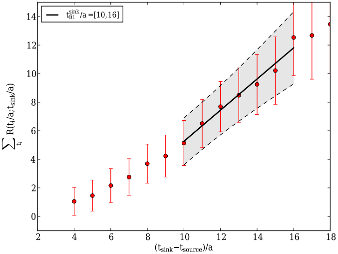




- **The one-end trick for the difference works very well suppressing the noise**
- **Clear rise of the plateau as t_S grows**
- **Disconnected piece around $\sim 10 - 15\%$ of the connected piece**

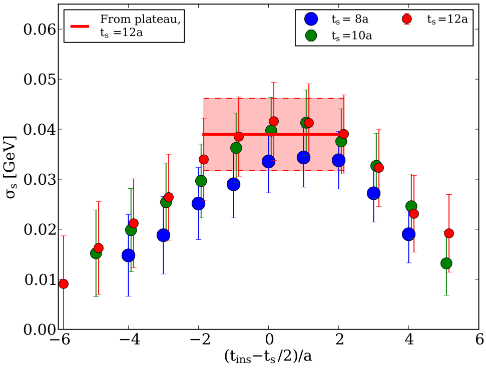
Results: Strange content of the nucleon





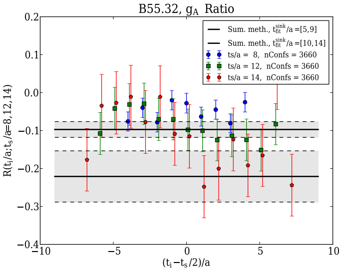
- Again, one-end trick for the difference
- Preliminary result, we will increase statistics
- The plateau seems to perfom better in this case

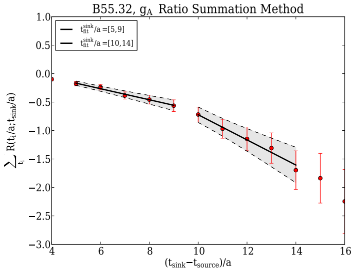
Results: $\sigma_{K\Delta}$ disconnected



- Preliminary result with small statistics, we need to investigate with more configurations if we have contamination here

Results: g_A disconnected





- **One-end trick for the sum doesn't provide μ noise suppression**
- **Consistent with former determinations of disconnected g_A, G. Bali et al. Phys.Rev.Lett.108 (2012), 222001**
- **Disconnected piece negative and $\sim 5 - 10\%$ of the connected piece**

Conclusions

- **GPU's suitable for computation of disconnected diagrams**
- **TSM highly reduces the variance while keeping the same computer cost**
- **The one-end trick for $u - d$ gives great results at low cost, but excludes flavour singlets**
- **The version for $u + d$ doesn't perform so well**
  - Noisy due to the presence of D_W or the lack of the noise suppression factor μ
  - Our current (partial) results for A_{20} and $\tilde{A}_{20}$ are inconclusive
- **The trick computes all time slices in a single inversion**
  - We can use both the plateau and the summation method
  - Plateau and summation methods give consistent results
- **It seems that some observables are affected by contamination coming from higher excitations**

Future plans

- **Improve the signal for all the successful observables using all the available momenta and statistics**
- **Compare the one-end trick for the sum with time dilution, taking into account other improvements**
  - Always in GPU's, to achieve maximum performance
- **Focus on flavour singlets**