Universität Bielefeld

Gitter-QCD und der Bielefelder GPU Cluster

Olaf Kaczmarek

Fakultät für Physik
Universität Bielefeld

German GPU Computing Group (G2CG)
Kaiserslautern
18.-19.04.2012

History of special purpose machines in Bielefeld

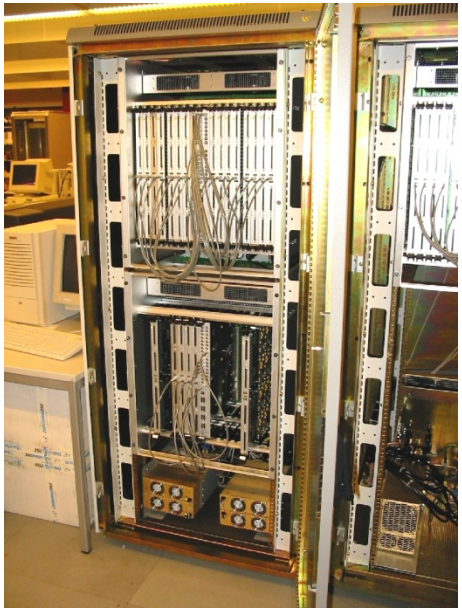
long history of dedicated lattice QCD machines in Bielefeld:

APE100 (procured 1993 – 1995)
25 GFlops peak

 # 4, 24, 25 hep-lat topcite all years

APE1000 (1999/2001)
144 GFlops

 # 17, 44, 50 hep-lat topcite all years



apeNEXT (2005/2006)
4000 GFlops

 # 3 2006 hep-lat topcite
 # 1 2007
 # 1 2008
 # 1 2009

Machines for Lattice QCD

Europe:

JUGENE in Jülich:
NIC-Project
PRACE-Project

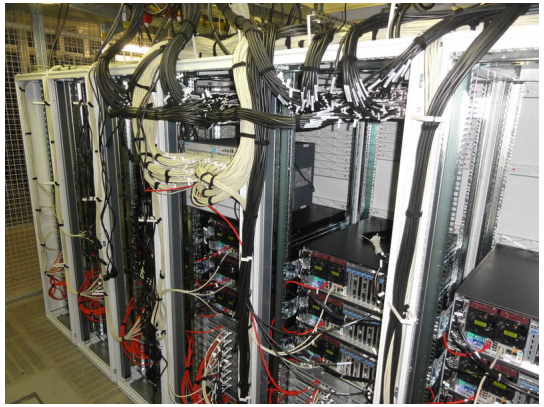
USA (USQCD):

+ Resources on New York Blue @ BNL
+ Bluegene/P in Livermore
+ GPU-Resources at Jefferson Lab

New GPU-Cluster of the Lattice Group in Bielefeld:

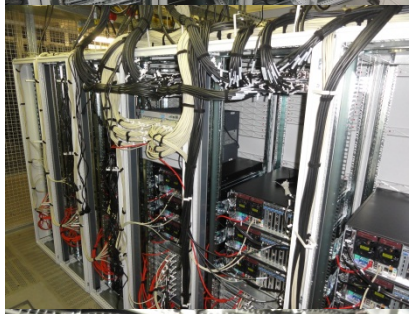
152 Nodes with 1216 CPU-Cores and 400 GPUs in total

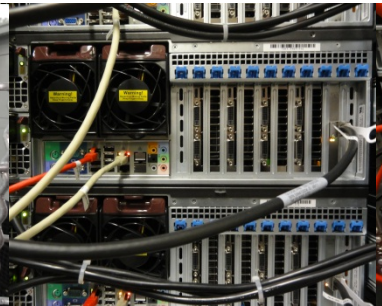
518 TFlops Peak Performance (single precision)
145 TFlops Peak Perfromance (double precision)



History of the new GPU cluster

Anfang 2009	Erste Gitter-QCD Portierung in CUDA
Anfang 2010	Konzeptausarbeitung + Vorbereitung des Antrags
10/10	Einreichung Großgeräteantrag
02/11	Rückfragen der Gutachter
07/11	Bewilligung
	Vorbereitung der Ausschreibung
09/11	Offene Ausschreibung
10/11	Zuschlag Fa. sysGen
11/11-01/12	Installation der Anlage
01/12	Einweihungsfeier
Anfang 2012	Start der ersten Physik Runs





Einweihung am 25.01.2012

Universität Bielefeld

Einweihung des neuen Bielefelder GPU-Clusters

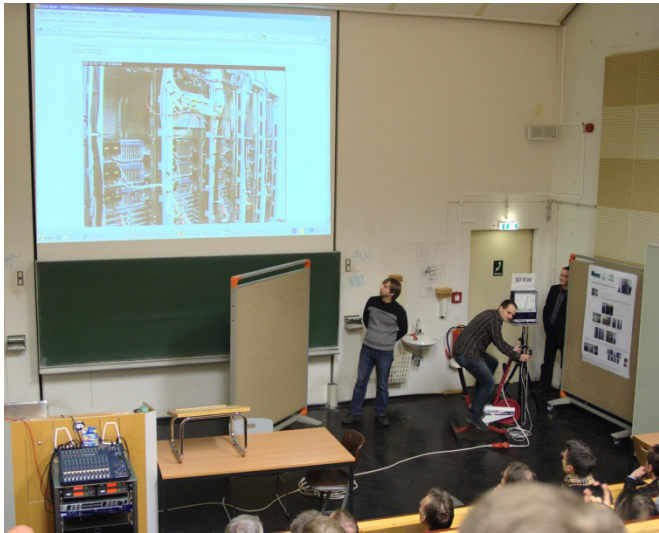
Grußworte

Prof. Martin Egelhaaf, *Prorektor Research, Bielefeld*
Prof. Andreas Hütten, *Dean Physics Dept., Bielefeld*

Prof. Peter Braun-Munzinger
(ExtreMe Matter Institute EMMI, GSI, TU Darmstadt and FIAS)
Nucleus-nucleus collisions at the LHC: from a gleam in the eye to quantitative investigations of the Quark-Gluon Plasma

Prof. Richard Brower
(Boston University)
QUDA: Lattice Theory on GPUs

Axel Köhler
(NVIDIA, Solution Architect HPC)
GPU Computing: Present and Future



Bielefeld GPU Cluster – Overview

Hybrid GPU HPC Cluster:

152 compute nodes

Number of GPUs: 400

Number of CPUs: 304 (1216 cores)

Total amount of CPU-memory: 7296 GB

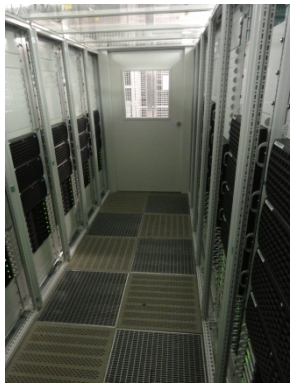
Total amount of GPU-memory: 1824 GB

14x19“ Racks incl. cold aisle containment

120-130kW Peak

< 10 kW/Rack

1x19“ Storage Server Rack



Peak performance:

CPUs:	**12 Tflops**
GPUs single precision:	**518 Tflops**
GPUs double precision:	**145 TFlops**

Bielefeld GPU Cluster – Compute Nodes

104 Tesla 1U-Knoten:

Dual Quadcore Intel Xeon CPU‘s

48 GB Memory

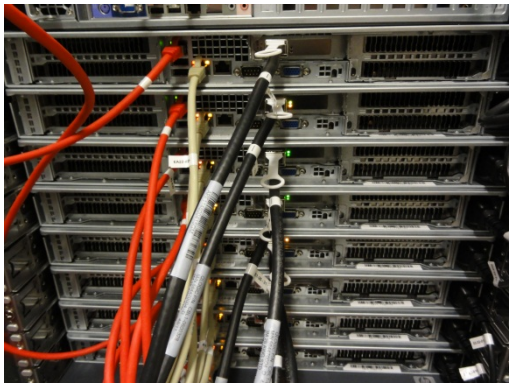
2x NVIDIA Tesla M2075-GPU (6GB ECC)

515 Gflops Peak double precision

1030 Gflops Peak single precision

150 GB/s Memory bandwidth

Total number of Tesla GPUs: 208



48 GTX580 4U-Knoten:

Dual Quadcore Intel Xeon CPU‘s

48 GB Memory

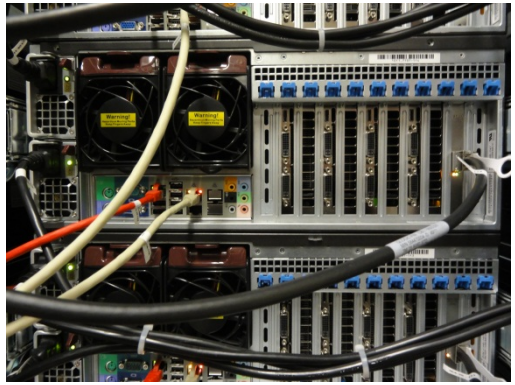
4x NVIDIA GTX580-GPU (3GB ECC)

198 Gflops Peak double precision

1581 Gflops Peak single precision

192 GB/s Memory bandwidth

Total number of GTX580 GPUs: 192



Bielefeld GPU Cluster – Compute Nodes

104 Tesla 1U-Knoten:

Dual Quadcore Intel Xeon CPU‘s

48 GB Memory

2x NVIDIA Tesla M2075-GPU (6GB ECC)

515 Gflops Peak double precision

1030 Gflops Peak single precision

150 GB/s Memory bandwidth

Total number of Tesla GPUs: 208

used for double precision calculations

+ when ECC error correction is important

48 GTX580 4U-Knoten:

Dual Quadcore Intel Xeon CPU‘s

48 GB Memory

4x NVIDIA GTX580-GPU (3GB ECC)

198 Gflops Peak double precision

1581 Gflops Peak single precision

192 GB/s Memory bandwidth

Total number of GTX580 GPUs: 192

used for fault tollerant measurements

+ when results can be checked

memory bandwidth still the limiting factor in Lattice QCD calculations, not performance

→ GTX580 faster even in double precision for most of our calculations

Bielefeld GPU Cluster – Head Nodes and Storage

Network:

QDR Infiniband network

(cluster nodes only x4-PCIe)

Gigabit network

IPMI remote management

2 Head Nodes:

Dual Quadcore Intel Xeon CPU‘s

48 GB Memory

Coupled as HA-Cluster

slurm queueing system

with GPUs as resources

and CPU jobs in parallel

7 Storage Nodes:

Dual Quadcore Intel Xeon CPU‘s

48 GB Memory

20 TB /home on 2-Server HA-Cluster

160 TB /work parallel filesystem

FhGFS distributed on 5 Servers

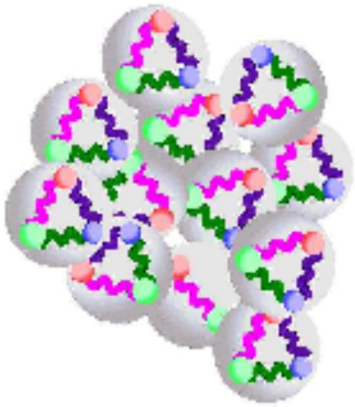
Infiniband connection to Nodes

3 TB Metadata on SSD

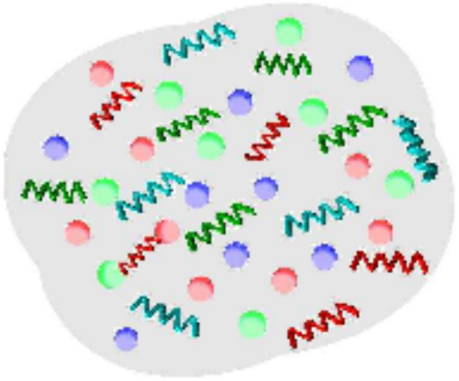
From Matter to the Quark Gluon Plasma

hadron gas

dense hadronic matter



quark gluon plasma

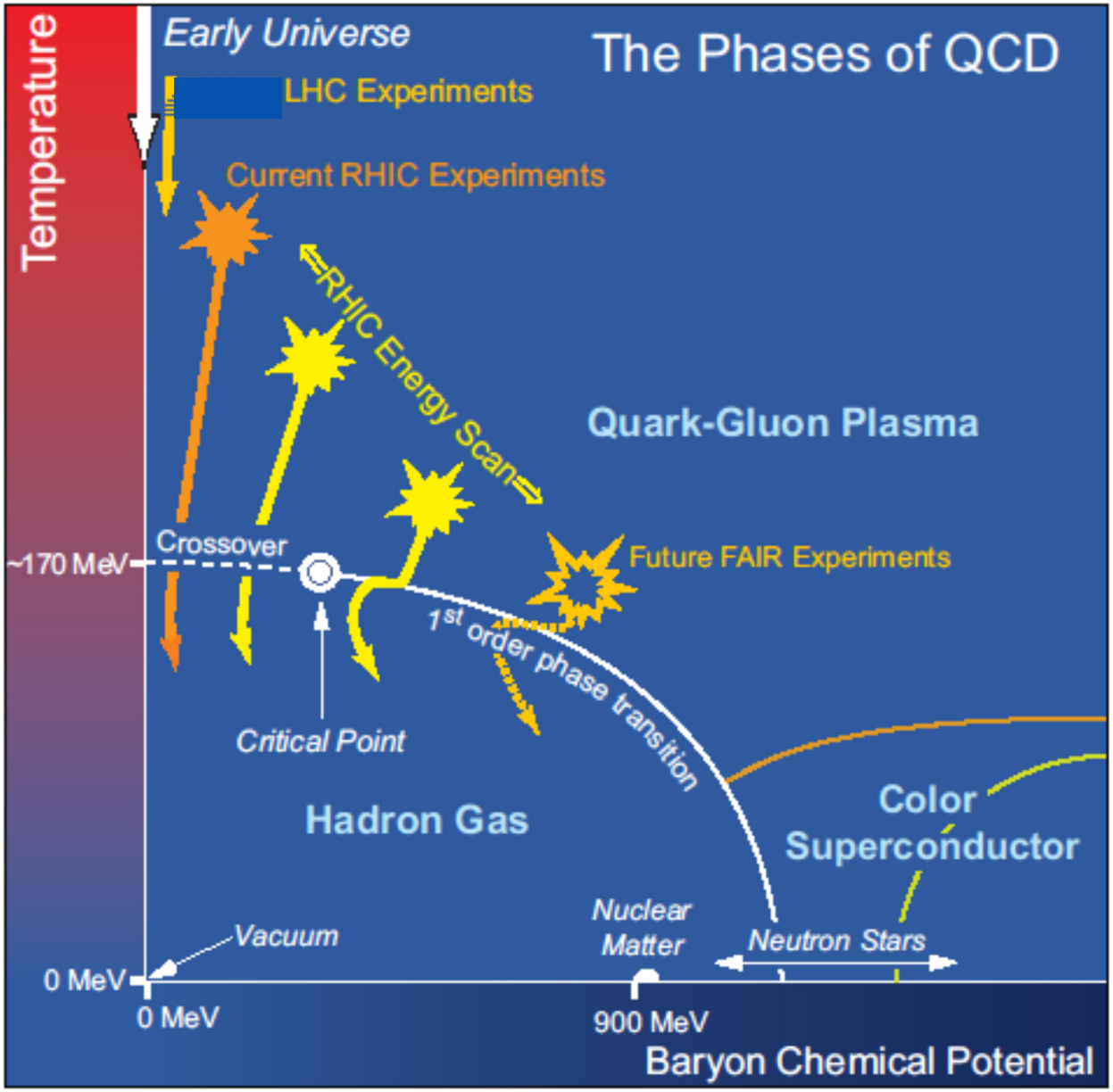


cold nuclear matter
Quarks and gluons are confined inside hadrons

phase transition or crossover at Tc

quarks and gluons are the degrees of freedom
(asymptotically) free

The Phases of Nuclear Matter



physics of the early universe:

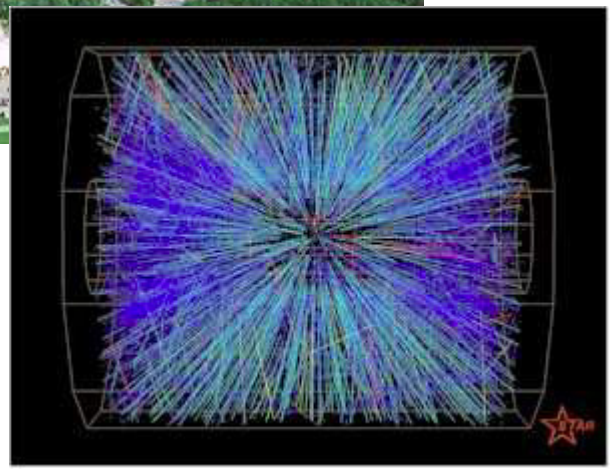
10^{-6} s after big bang

very hot: $T \sim 10^{12}$ K

experimentally accessible in Heavy Ion Collisions at SPS, RHIC, LHC, FAIR

very dense: $n_B \sim 10\ n_{NM}$

Heavy Ion Experiments – RHIC@BNL

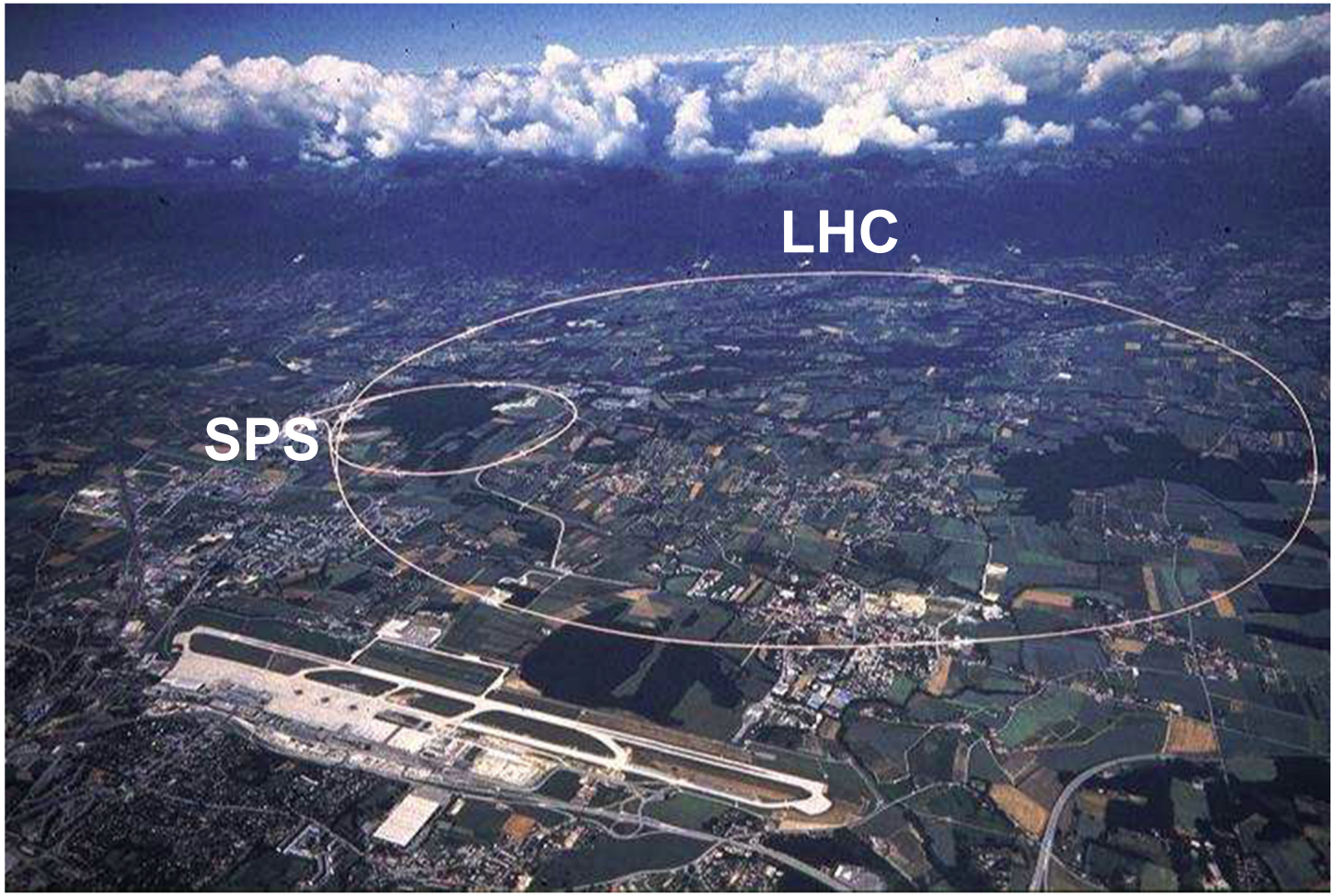


Au-Au beams with $\sqrt{s}$ = 130, 200 GeV/A

estimated initial temperature: $T_0 \simeq (1.5\text{-}2)\, T_c$

estimated initial energy density: $\varepsilon_0 \simeq (5\text{-}15)$ GeV/fm^3

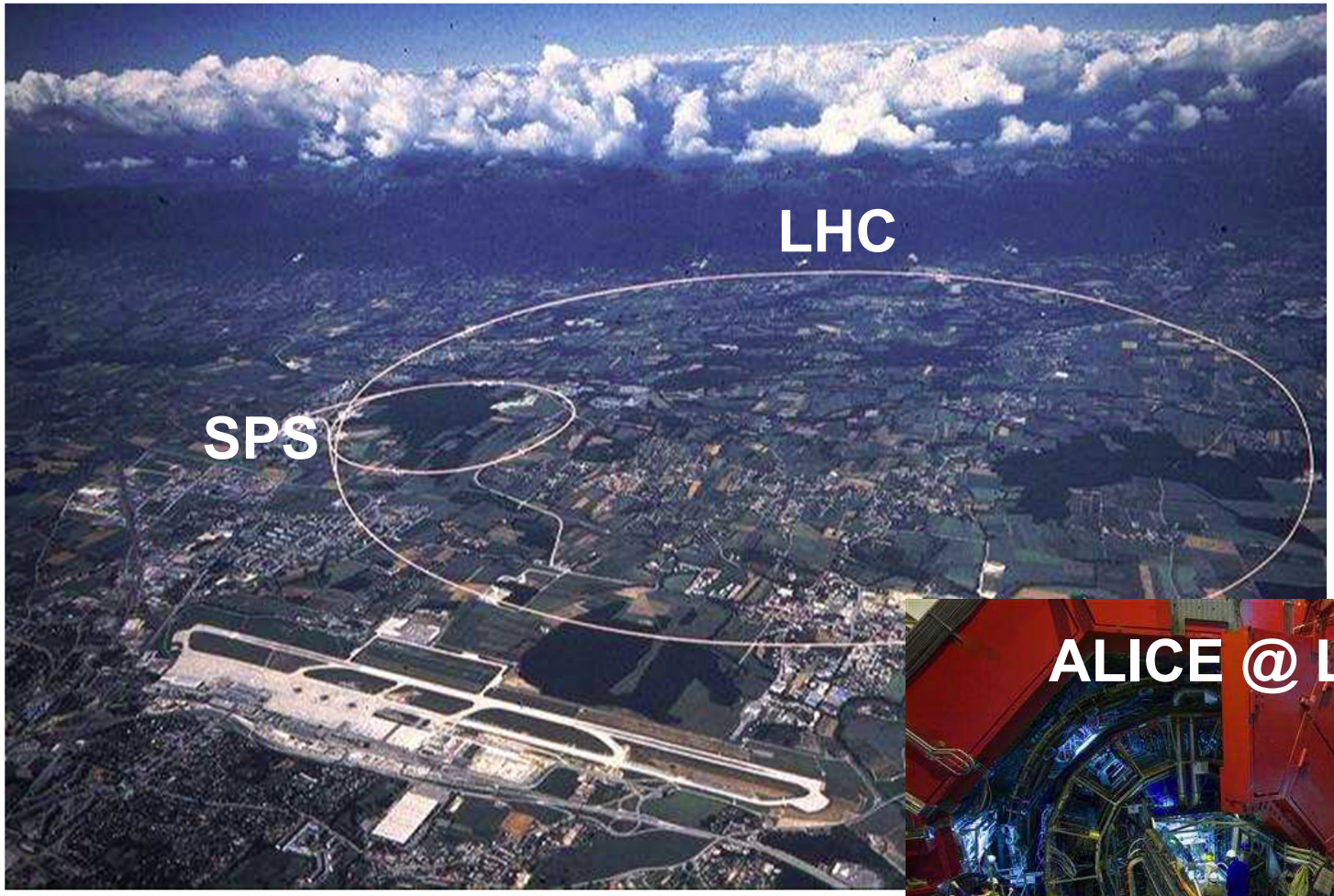
Heavy Ion Experiments – LHC@CERN



Pb-Pb beams with $\sqrt{s}$ = 2.7 TeV/A

estimated initial temperature: $T_0 \simeq (2\text{-}3)\, T_c$

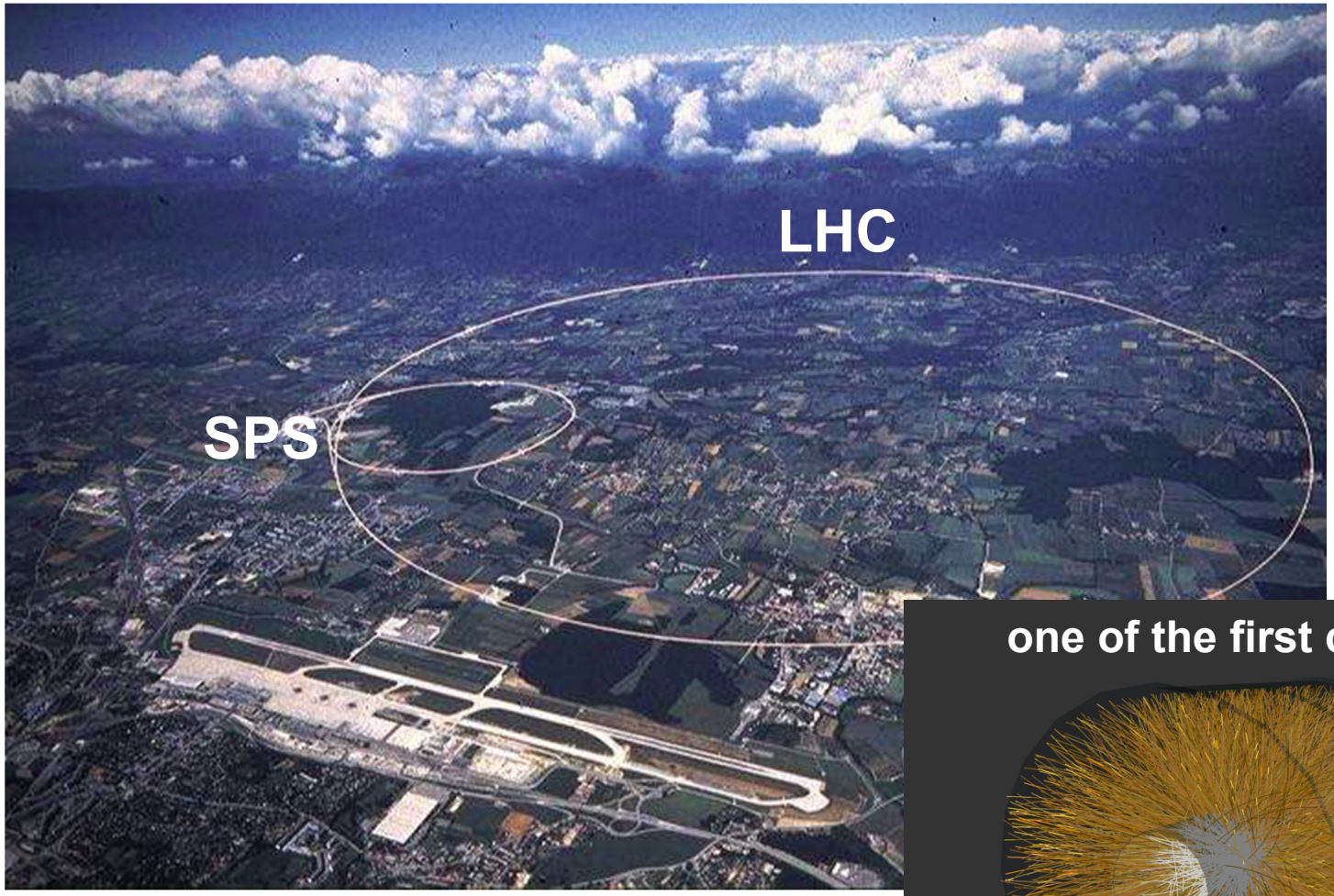
Heavy Ion Experiments – LHC@CERN

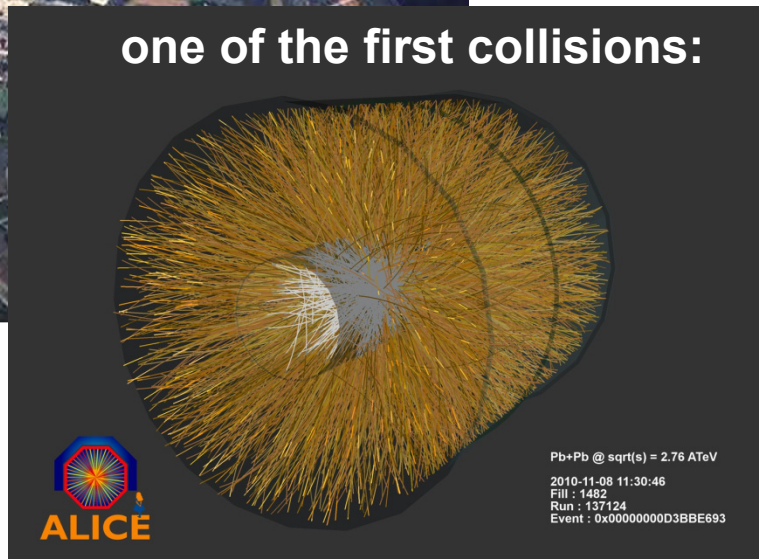


Pb-Pb beams with $\sqrt{s}$ = 2.7 TeV/A

estimated initial temperature: $T_0 \simeq (2\text{-}3)\, T_c$

Heavy Ion Experiments – LHC@CERN

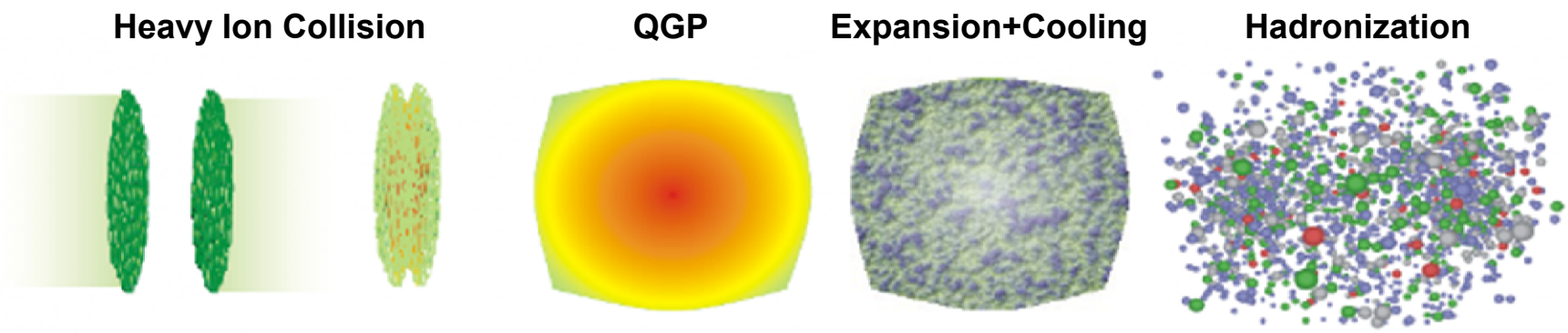




Pb-Pb beams with $\sqrt{s}$ = 2.7 TeV/A

estimated initial temperature: $T_0 \simeq (2\text{-}3)\, T_c$

Evolution of Matter in a Heavy Ion Collisions



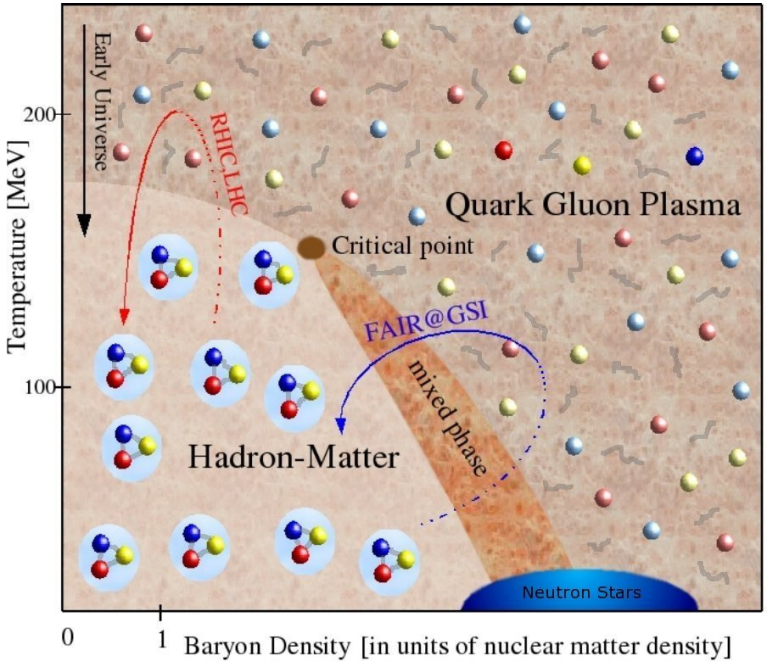
detectors only measure particles after hadronization

→ need to understand the whole evolution of the system

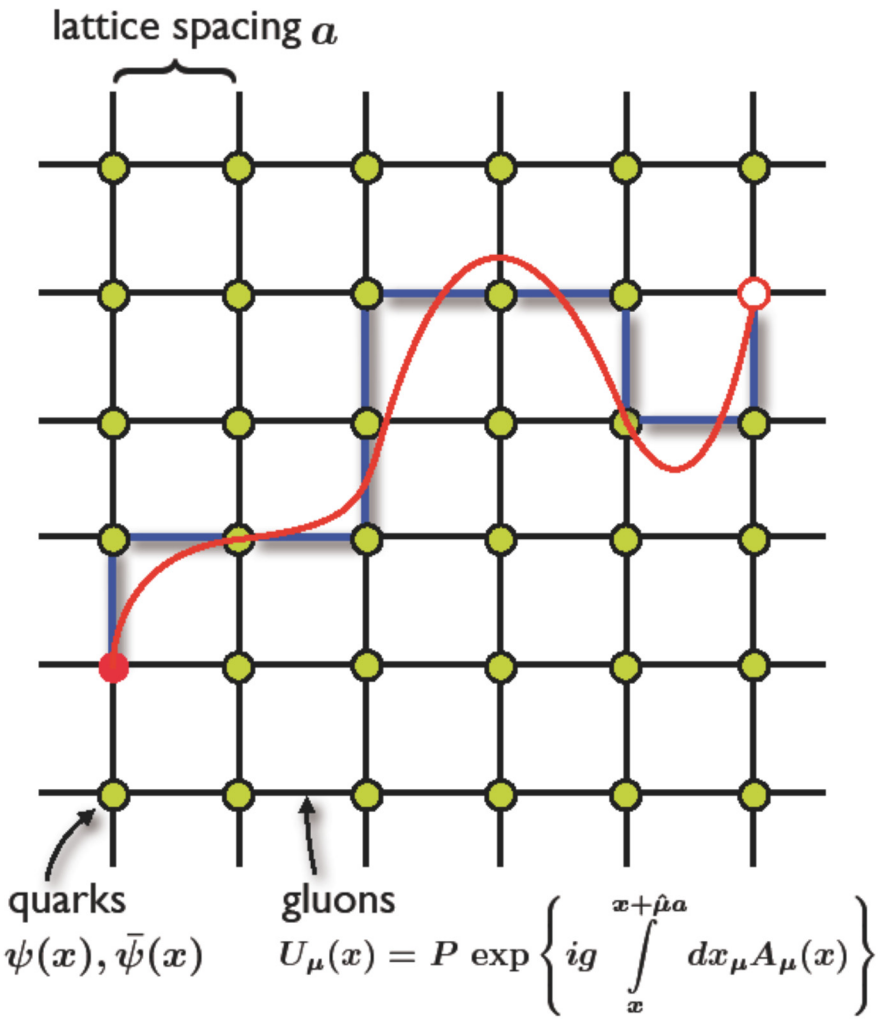
→ theoretical input from ab initio non-perturbative calculations

→ equation of state, critical temperature, pressure, energy, fluctuations, critical point, ...

→ Lattice QCD



Lattice QCD – Discretization of space/time



Gluons: $U_\mu(x) \in SU(3)$

complex 3x3 matrix per link

18 (12/8) float per link

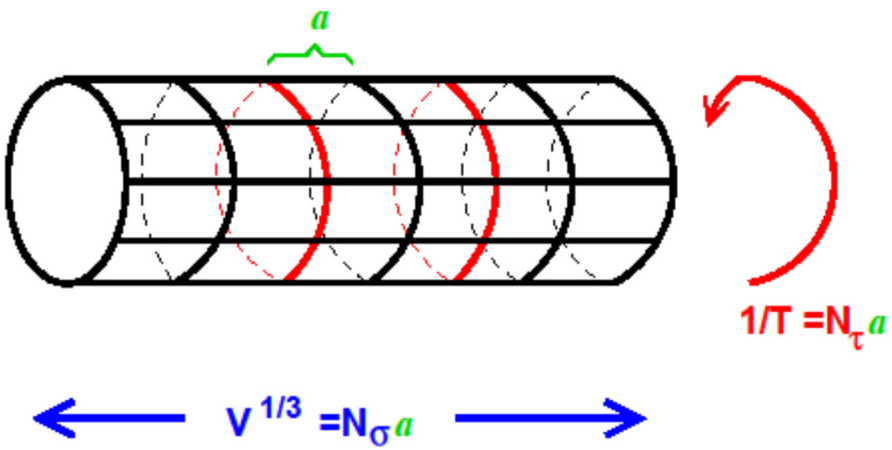
Quarks: Fermion-fields described by Grassmann variables

$$\begin{aligned} \psi_1 \psi_2 &= -\psi_2 \psi_1 \\ \psi^2 &= 0 \end{aligned}$$

Calculations at finite lattice spacing a and finite volume $N_s^3 \times N_t^3$

Thermodynamic limit: $V \longrightarrow \infty$

Continuum limit: $a \longrightarrow 0$



Quantum Chromo Dynamics at finite Temperature

partition function:

$$Z(T,V,\mu) = \int \mathcal{D}\mathcal{U}\ \mathcal{D}\psi\mathcal{D}\psi e^{-S_E[U,\psi,\psi]}$$

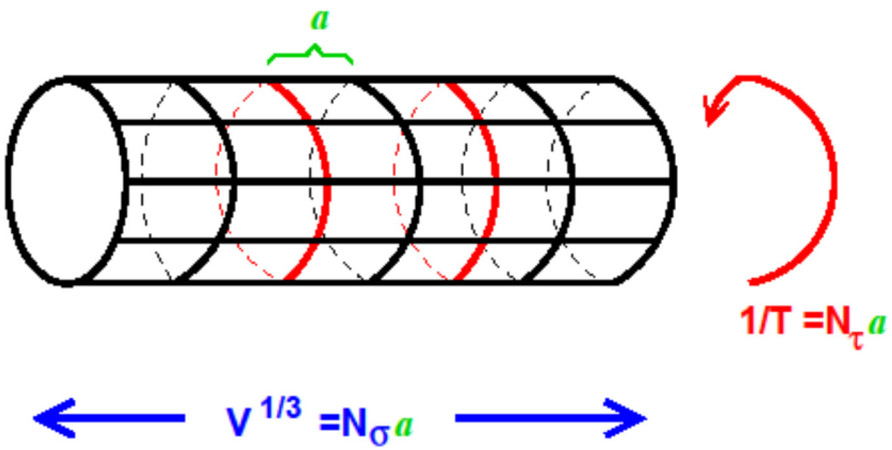
$$\boxed{S_E = \int^{1/T} d_{x_0} \int_V d^3x \mathcal{L}_E(U,\psi,\psi,\mu)}$$

temperature volume

Calculations at finite lattice spacing a and finite volume

Thermodynamic limit: $V \longrightarrow \infty$

Continuum limit: $a \longrightarrow 0$



Quantum Chromo Dynamics at finite Temperature

partition function:

$$Z(T,V,\mu) = \int \mathcal{DU}\ \mathcal{D}\psi\mathcal{D}\psi e^{-S_E[U,\psi,\psi]}$$

$$S_E = \int^{1/T} d_{x_0} \int_V d^3x \mathcal{L}_E(U,\psi,\psi,\mu)$$

temperature volume

Hybrid Monte Carlo calculations:

generate gauge fields U with probability

$$P[U] = \frac{1}{Z} e^{-S_E}$$

using molecular dynamics evalaluation

in a fictitious time in configuration space

(Markov Chain [U_1], [U_2], ...)

The QCD partition function

- the QCD partition function:

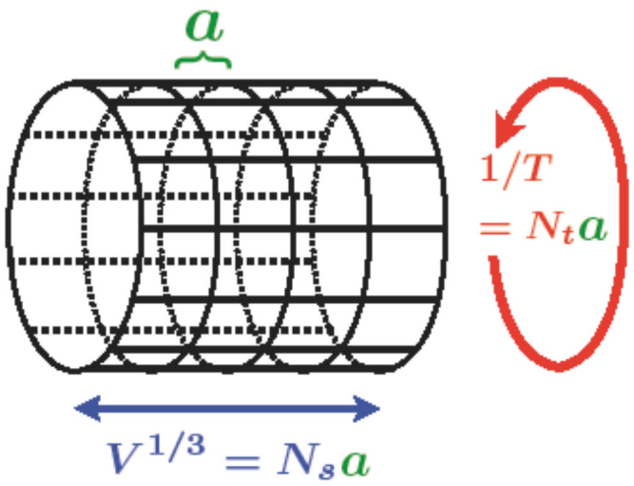
$$Z(V,T,\bar{\mu}) = \int \mathcal{D}A\, \mathcal{D}\bar{\psi}\, \mathcal{D}\psi\ \exp\{-S_E\}$$

$$S_E = \bar{\psi}_x M_{x,y} \psi_y + S_G$$

$$M_{x,y} = am\,\delta_{x,y} + \frac{1}{2}\sum_{\mu=1}^{3}\gamma_\mu\left\{U_\mu(x)\,\delta_{x+a\hat{\mu},y} - U_\mu^\dagger(y)\,\delta_{x-a\hat{\mu},y}\right\}$$

$$+\frac{1}{2}\gamma_4\left\{e^{a\bar{\mu}}\,U_4(x)\,\delta_{x+a\hat{4},y} - e^{-a\bar{\mu}}\,U_4^\dagger(y)\,\delta_{x-a\hat{4},y}\right\}$$

- geometry of space time: $N_s^3 \times N_t$ (4d - torus)



note:

- only closed loops participate to the partition function
- only loops that wind around the torus in time direction $\mathcal{W}$-times pick up a μ-dependence:

$$\exp\{\mathcal{W}\mu/T\}$$

$\longrightarrow$ alternatively (gauge-transformation):

- choose a fixed time-slice on which all temporal links get a factor $\exp\{\pm\mu/T\}$

The QCD partition function

- integration over fermion fields

$$Z(V,T,\mu) = \int \mathcal{D}A\mathcal{D}\psi\mathcal{D}\bar{\psi}\ \exp\{S_F(A,\psi,\bar{\psi}) - \beta S_G(A)\}$$

$$= \int \mathcal{D}A\ \det[M](A,\mu)\exp\{-\beta S_G(A)\}$$

complex for $\mu > 0$

propabilistic interpretation necessary for Monte Carlo!

Taylor expansion at finite density

- Taylor expansion:
  - start from Taylor expansion of the pressure,

$$\frac{p}{T^4} = \frac{1}{VT^3} \ln Z(V,T,\mu_u,\mu_d,\mu_s) = \sum_{i,j,k} c^{u,d,s}_{i,j,k} \left(\frac{\mu_u}{T}\right)^i \left(\frac{\mu_d}{T}\right)^j \left(\frac{\mu_s}{T}\right)^k$$

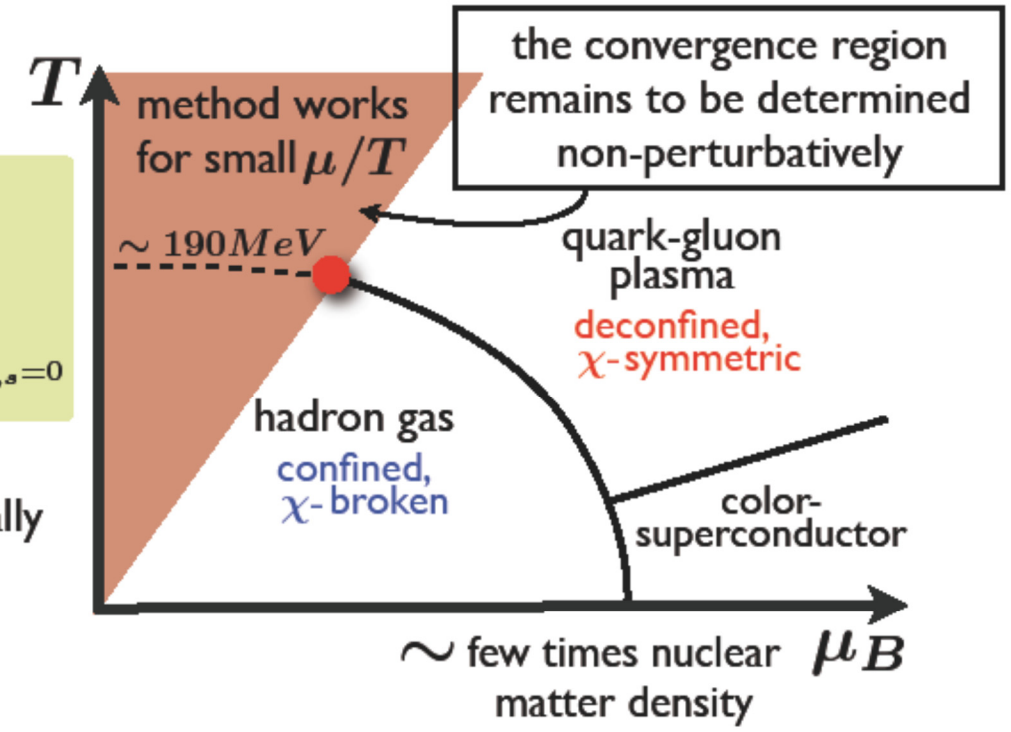
  - calculate expansion coefficients for fixed temperature

  - no sign problem:
    all simulations are done at $\mu = 0$

$$c^{u,d,s}_{i,j,k} \equiv \frac{1}{i!j!k!} \frac{1}{VT^3} \cdot \frac{\partial^i \partial^j \partial^k \ln Z}{\partial(\frac{\mu_u}{T})^i \partial(\frac{\mu_d}{T})^j \partial(\frac{\mu_s}{T})^k}\Bigg|_{\mu_{u,d,s}=0}$$

  - method is straight forward:
    all terms can be generated automatically

Allton *et al.*, PRD66:074507,2002;
Allton *et al.*, PRD68:014507,2003;
Allton *et al.*, PRD71:054508,2005.
(see also publications by
MILC and Gavai, Gupta)



Taylor expansion at finite density

- formulate all operators in term of space-time, color (and spin) traces:

$$
\begin{aligned}
\frac{\partial(\ln\det M)}{\partial\mu} &= \mathcal{D}_1 &&= \mathrm{Tr}\left(M^{-1}\frac{\partial M}{\partial\mu}\right)\\
\frac{\partial^2(\ln\det M)}{\partial\mu^2} &= \mathcal{D}_2 &&= \mathrm{Tr}\left(M^{-1}\frac{\partial^2 M}{\partial\mu^2}\right) - \mathrm{Tr}\left(M^{-1}\frac{\partial M}{\partial\mu}M^{-1}\frac{\partial M}{\partial\mu}\right)\\
\frac{\partial^3(\ln\det M)}{\partial\mu^3} &= \mathcal{D}_3 &&= \mathrm{Tr}\left(M^{-1}\frac{\partial^3 M}{\partial\mu^3}\right) - 3\mathrm{Tr}\left(M^{-1}\frac{\partial M}{\partial\mu}M^{-1}\frac{\partial^2 M}{\partial\mu^2}\right)\\
&&&\quad +2\mathrm{Tr}\left(M^{-1}\frac{\partial M}{\partial\mu}M^{-1}\frac{\partial M}{\partial\mu}M^{-1}\frac{\partial M}{\partial\mu}\right)\\
\frac{\partial^4(\ln\det M)}{\partial\mu^4} &= \mathcal{D}_4 &&= \mathrm{Tr}\left(M^{-1}\frac{\partial^4 M}{\partial\mu^4}\right) - 4\mathrm{Tr}\left(M^{-1}\frac{\partial M}{\partial\mu}M^{-1}\frac{\partial^3 M}{\partial\mu^3}\right)\\
&&&\quad -3\mathrm{Tr}\left(M^{-1}\frac{\partial^2 M}{\partial\mu^2}M^{-1}\frac{\partial^2 M}{\partial\mu^2}\right) + 12\mathrm{Tr}\left(M^{-1}\frac{\partial M}{\partial\mu}M^{-1}\frac{\partial M}{\partial\mu}M^{-1}\frac{\partial^2 M}{\partial\mu^2}\right)\\
&&&\quad -6\mathrm{Tr}\left(M^{-1}\frac{\partial M}{\partial\mu}M^{-1}\frac{\partial M}{\partial\mu}M^{-1}\frac{\partial M}{\partial\mu}M^{-1}\frac{\partial M}{\partial\mu}\right)
\end{aligned}
$$

- evaluate all traces by noisy estimators:

$$
\mathrm{Tr}\left(\frac{\partial^{n_1} M}{\partial\mu^{n_1}}M^{-1}\frac{\partial^{n_2} M}{\partial\mu^{n_2}}\cdots M^{-1}\right) = \lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}\eta_k^\dagger\frac{\partial^{n_1} M}{\partial\mu^{n_1}}M^{-1}\frac{\partial^{n_2} M}{\partial\mu^{n_2}}\cdots M^{-1}\eta_k
$$

with N random vectors, satisfying $\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\eta^*_{n,i}\eta_{n,j} = \delta_{i,j}$

- construct expansion coefficients from $\mathcal{D}_n^u, \mathcal{D}_n^d, \mathcal{D}_n^s$, with unbiased estimators

$$
c_{2,0,0}^{u,d,s} = \frac{1}{2}\frac{N_\tau}{N_\sigma^3}\left(\langle\mathcal{D}_2^u\rangle + \left\langle\left(\mathcal{D}_1^u\right)^2\right\rangle\right)
$$

Matrix inversion

Iterative solvers, e.g. Conjugate Gradient: $M(U)\chi = \psi$

sparse matrix M

only non-zero elements $U_\mu(x)$ are stored

each thread calculates one lattice point x

typical CUDA kernel for M$\times$v multiplication:

```
for(mu=0; mu<4; mu++)
        for(nu=0; nu<4; nu++)
            if(mu!=nu)
            {
              site_3link = GPUsu3lattice_indexUp2Up(xl, yl, zl, tl, mu, nu, c_latticeSize );
              x_1 = x+c_latticeSize.vol4()*munu+(12*c_latticeSize.vol4())*0;
              v_2[threadIdx.x] += g_u3.getElement(x_1) * g_v.getElement(site_3link-c_latticeSize.sizeh());

              site_3link = GPUsu3lattice_indexDown2Up(xl, yl, zl, tl, mu, nu, c_latticeSize );
              x_1 = site_3link+c_latticeSize.vol4()*munu+(12*c_latticeSize.vol4())*1;
              v_2[threadIdx.x] -= tilde(g_u3.getElement(x_1)) * g_v.getElement(site_3link-c_latticeSize.sizeh());

              site_3link = GPUsu3lattice_indexUp2Down(xl, yl, zl, tl, mu, nu, c_latticeSize );
              x_1 = x+c_latticeSize.vol4()*munu+(12*c_latticeSize.vol4())*1;
              v_2[threadIdx.x] += g_u3.getElement(x_1) * g_v.getElement(site_3link-c_latticeSize.sizeh());

              site_3link = GPUsu3lattice_indexDown2Down(xl, yl, zl, tl, mu, nu, c_latticeSize );
              x_1 = site_3link+c_latticeSize.vol4()*munu+(12*c_latticeSize.vol4())*0;
              v_2[threadIdx.x] -= tilde(g_u3.getElement(x_1)) * g_v.getElement(site_3link-c_latticeSize.sizeh());

                  munu++;
            }
```

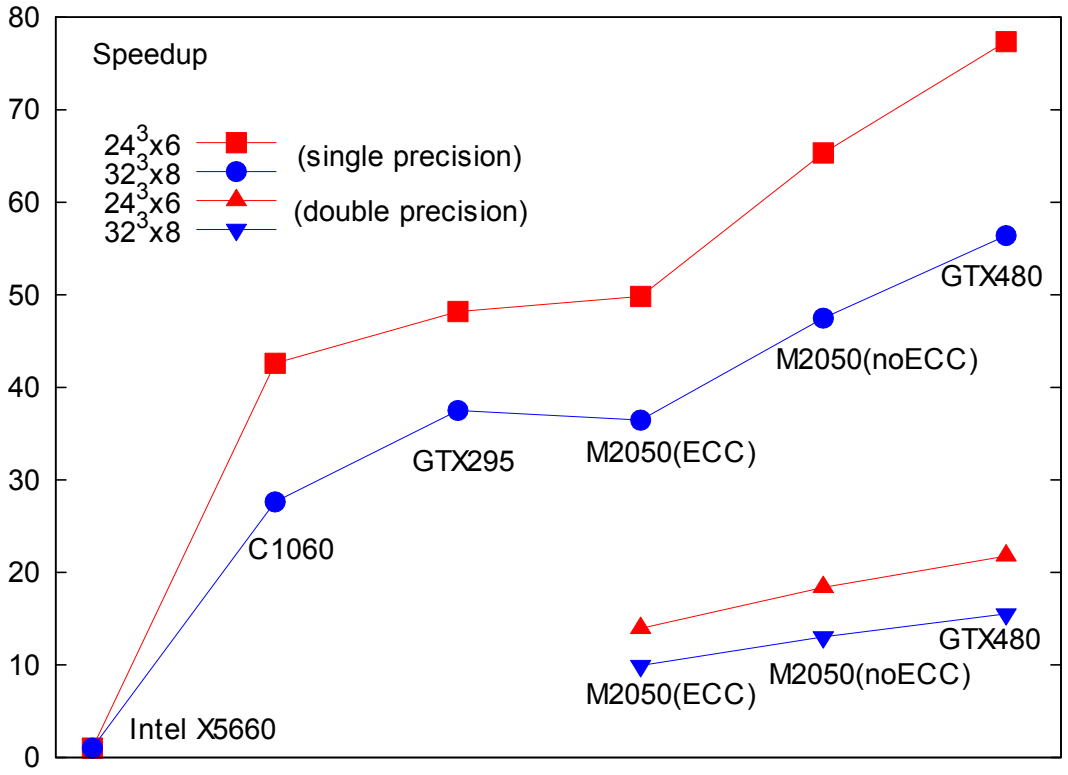
Performance for matrix inverter

typical lattice sizes:

$32^3 \times 8$ 72MB (single) 144MB (double)

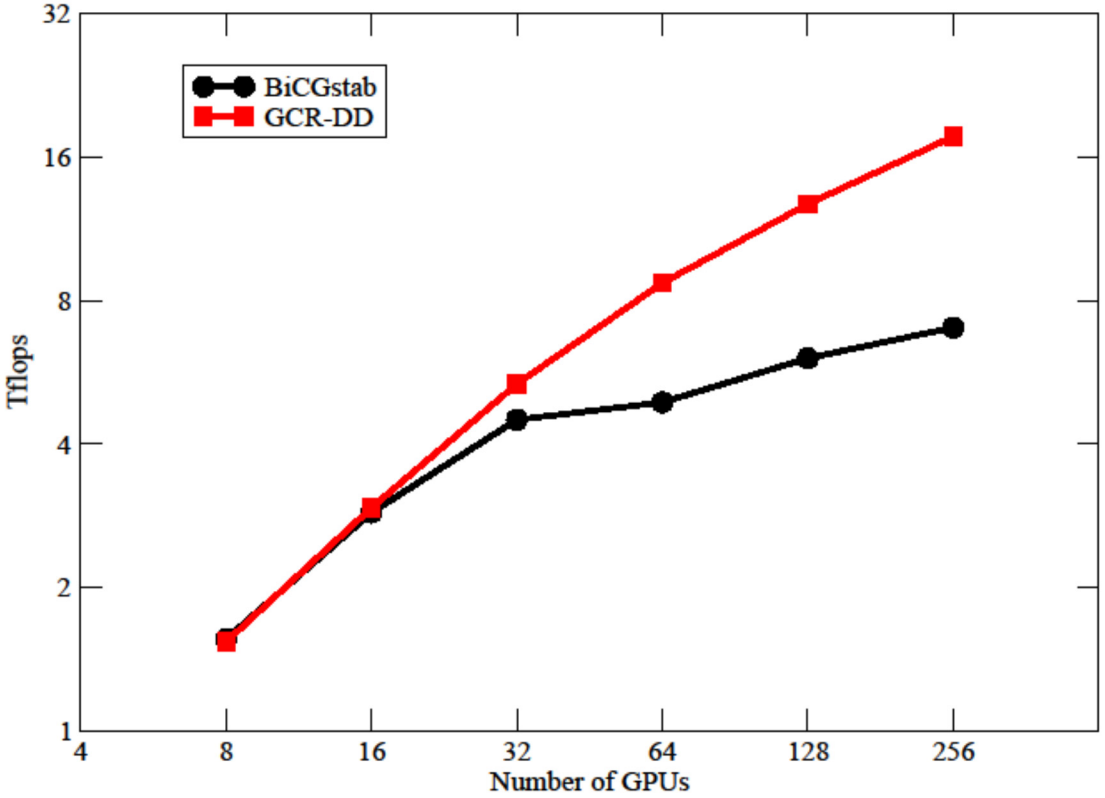
$48^3 \times 12$ 365MB (single) 730MB (double)

so far only single-GPU code



Multi-GPU matrix inverter

Scaling Lattice QCD beyond 100 GPUs, R.Babich, M.Clark et al., 2011



Bielefeld – BNL Collaboration

Most work is done by people, not by machine:

Bielefeld:

Frithjof Karsch

Edwin Laermann
Olaf Kaczmarek
Markus Klappenbach

Mathias Wagner
Christian Schmidt
Dominik Smith

Marcel Müller

Thomas Luthe
Lukas Wresch

Regensburg:

Wolfgang Soeldner

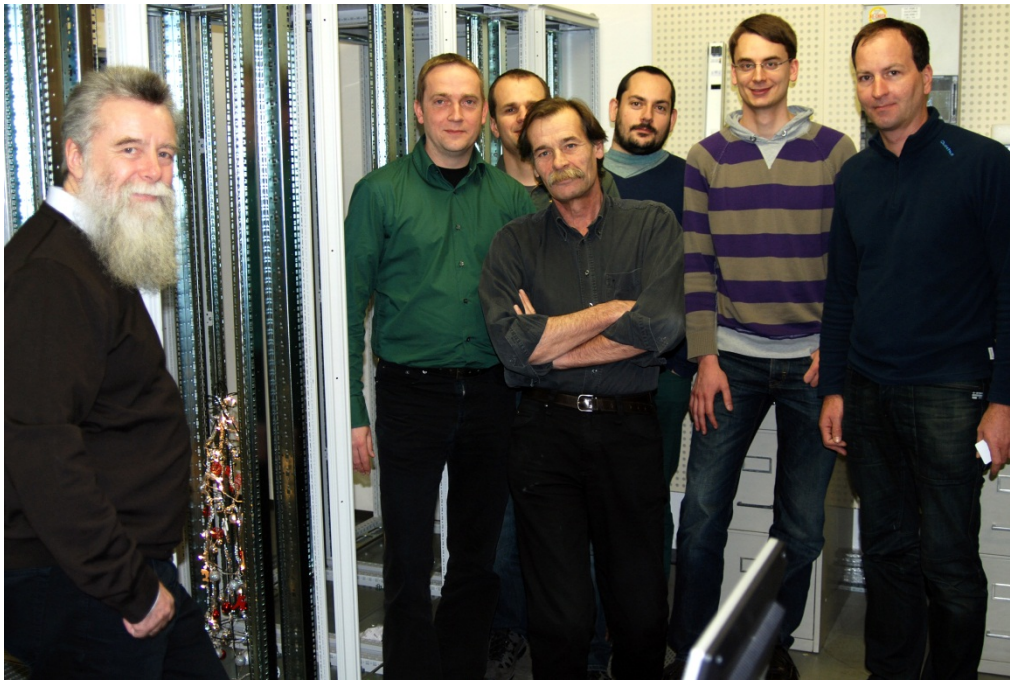
Krakow:

Piotr Bialas

Brookhaven National Lab:

Peter Petreczky
Swagato Mukherjee

Aleksy Bazavov
Heng-Tong Ding
Prasad Hegde
Yu Maezawa

+ a lot of help from:

M.Clark (NVIDIA QCD-Team)
and
M.Bach (FIAS Frankfurt)