HQS@HPC
Improving Scalability of large sparse ED studies on HLRB-II

G. Wellein, G. Hager

HPC Services

Regionales Rechenzentrum Erlangen

Friedrich-Alexander-University Erlangen-Nuremberg

Germany

H. Fehske, A. Alvermann

Chair for Complex Quantum Systems

Institute of Physics

University of Greifswald

Germany

KONWIHR Review Workshop - July 2007

Survey

- **Motivation**
- **Improving Scalability of sparse ED code for HLRB-II**
- **Current progress in physics**

Motivation: From Physics to Supercomputers

Physics

Microscopic Model: Hamiltonian ***H***

Ground state properties

Thermodynamics excitation spectra

Dynamical correlation functions

Highly correlated electron/spin-phonon Systems

High-T_c Cuprates, CMR manganites, Quasi-1D Metal/Spin Systems

„The whole is greater than its parts“

Exact numerical studies of finite systems using Supercomputers

Motivation: Numerical Methods of HQS@HPC

- **Density Matrix Renormalization Group (DRMG) Method**
  - Originally introduced by White in 1992
  - Very efficient for ground-state properties in (quasi) 1D models
  - Efficient parallel implementation (4-16 CPUs) through KONWIHR support (2003-2004)

PHYSICAL REVIEW B **71**, 075108 (2005)

Stripe formation in doped Hubbard ladders

G. Hager and G. Wellein
Regionales Rechenzentrum Erlangen, Martensstraße 1, D-91058, Erlangen, Germany

E. Jeckelmann
Institut für Physik, Johannes Gutenberg-Universität, D-55099 Mainz, Germany

H. Fehske
Institut für Physik, Ernst-Moritz-Arndt-Universität, Greifswald, D-17489 Greifswald, Germany

(Received 13 September 2004; published 10 February 2005)

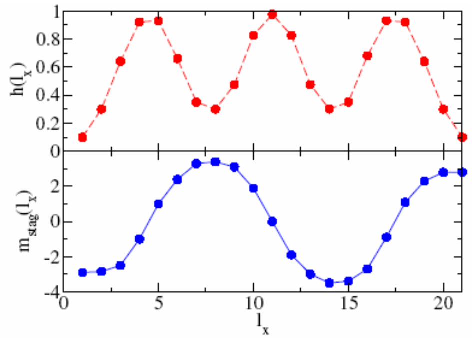


FIG. 2 (color online). This figure shows DMRG results (G. Hager *et al.* [10]) for the hole $h(\ell_x)$ (dashed red line) and the staggered spin $m_{stag}(\ell_x)$ (solid blue line) densities along the leg direction for a 21×6 Hubbard ladder with 12 holes and $U/t = 12$. As discussed in the text, $h(\ell_x)$ corresponds to the spin polarization $n_s(\ell_x)$ and $m_{stag}(\ell_x)$ corresponds to the s-wave pairfield order parameter $\Delta(\ell_x)$ of the FFLO state.

Selected references to this paper:

Moreo A, Scalapino DJ
PHYSICAL REVIEW LETTERS 98 (21): Art. No. 216402 MAY 25 2007

Feiguin AE, White SR, Scalapino DJ
PHYSICAL REVIEW B 75 (2): Art. No. 024505 JAN 2007

Motivation: Numerical Methods of HQS@HPC

- **Exact Diagonalization (ED)**
  - Physical parameter space restricted by available computer resources
  - Approximation free – 2D/3D & excitation spectra
  - Choose complete basis set -> sparse matrix problem -> Increase matrix as far as possible
  - First ED studies of correlated electron-phonon systems: 1992/93
  - Finite temperature & CPT integration (KONWIHR 2005)

  - Focus of 2006 KONWIHR activities: Scalability issues on HLRB-II

Motivation: Sparse ED implementation

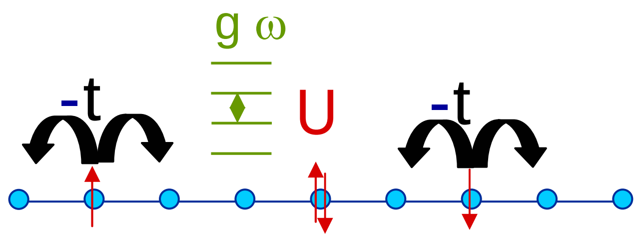
- Time & memory consuming step: Sparse Matrix Vector Multiplication
- Out-of-core implementation (do not store non-zero elements)
- Largest ED study known so far: Matrix dimension $D_{max}= 1.59*10^{11}$ (Yamada et al., SC2005)

System	CM5 GMD/ St. Augustin	CRAY T3E NIC Jülich	HLRB-I (Hitachi SR8k) LRZ Munich	**HLRB-II (SGI Altix) LRZ Munich**
Production	1993/1994	1998-2001	2001-2005	**2006-**
#CPU Memory	64 2 GB	256 128 GB	1216 900 GB	**5720 (cores)** **16,000 GB**
Parallelis.	CMFortran	MPI/SHMEM	MPI+OpenMP	**MPI(+shmem)**
$\mathbf{D_{max}}$	$\mathbf{5.6 * 10^{7}}$	$\mathbf{4.4 * 10^{9}}$	$\mathbf{3.3 * 10^{10}}$	$\mathbf{3.8 * 10^{11}}$
MVM [s]	**156**	**33**	**63**	**38**

Improving Scalability:

Parallelism in Holstein(-Hubbard) type models

$$H = -t \sum_{\langle ij\rangle,\sigma} \left[c^\dagger_{i\sigma} c_{j\sigma} + \mathrm{H.c.}\right] + U \sum_i n_{i\uparrow} n_{i\downarrow} + g\omega_0 \sum_{i,\sigma} (b_i^\dagger + b_i) n_{i\sigma} + \omega_0 \sum_i b_i^\dagger b_i$$



- Hilbert space: Direct product of electronic & phononic Hilbert space:
  `{|e>|p>; e=1,…,D`$_{el}$`; p=1,…, D`$_{ph}$`}`
- A vector is defined as
  `|v> = Σ v`$_{e,p}$` [|e>|p>]`
- Distribute "electronic part" to `n`$_{pro}$ processes (`rank=0,…,n`$_{pro}$`-1`):
  `{v`$_{e,p}$`; e=(rank*(D`$_{el}$`/n`$_{pro}$`)+1,…,(rank+1)*(D`$_{el}$`/n`$_{pro}$`); p=1,…, D`$_{ph}$`}`
- Impact on matrix vector multiplication:
  - Communication may be generated by hopping (t) term
  - Contribution of phonon operators can be computed locally
  - Number of parallel processes is limited by D_{el}

Improving Scalability:
Original implementation

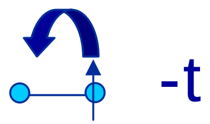
Holstein model: 2 sites, 1 electron (D_{el}=2) and D_{ph} phononic states

Running hopping part of matrix vector multiplication in parallel on 2 processors:

{ |↑,0> |p>; p=1,…, D_{ph}}

Program flow `rank=0`

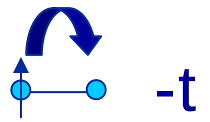
```
real*8 new(D_ph,1), old(D_ph,1)
real*8 rbuf(D_ph)
…
call shmem_get(rbuf,old, dest=1)

do i=1,Dph
  new(i,1) = new(i,1) -t*rbuf(i)
enddo
…
```



{ |0,↑> |p>; p=1,…, D_{ph}}

Program flow `rank=1`

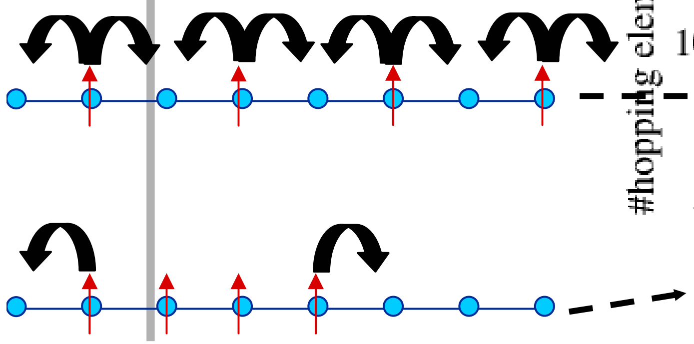
```
real*8 new(D_ph,1), old(D_ph,1)
real*8 rbuf(D_ph)
…
call shmem_get(rbuf,old, dest=0)

do i=1,Dph
  new(i,1) = new(i,1) -t*rbuf(i)
enddo
…
```

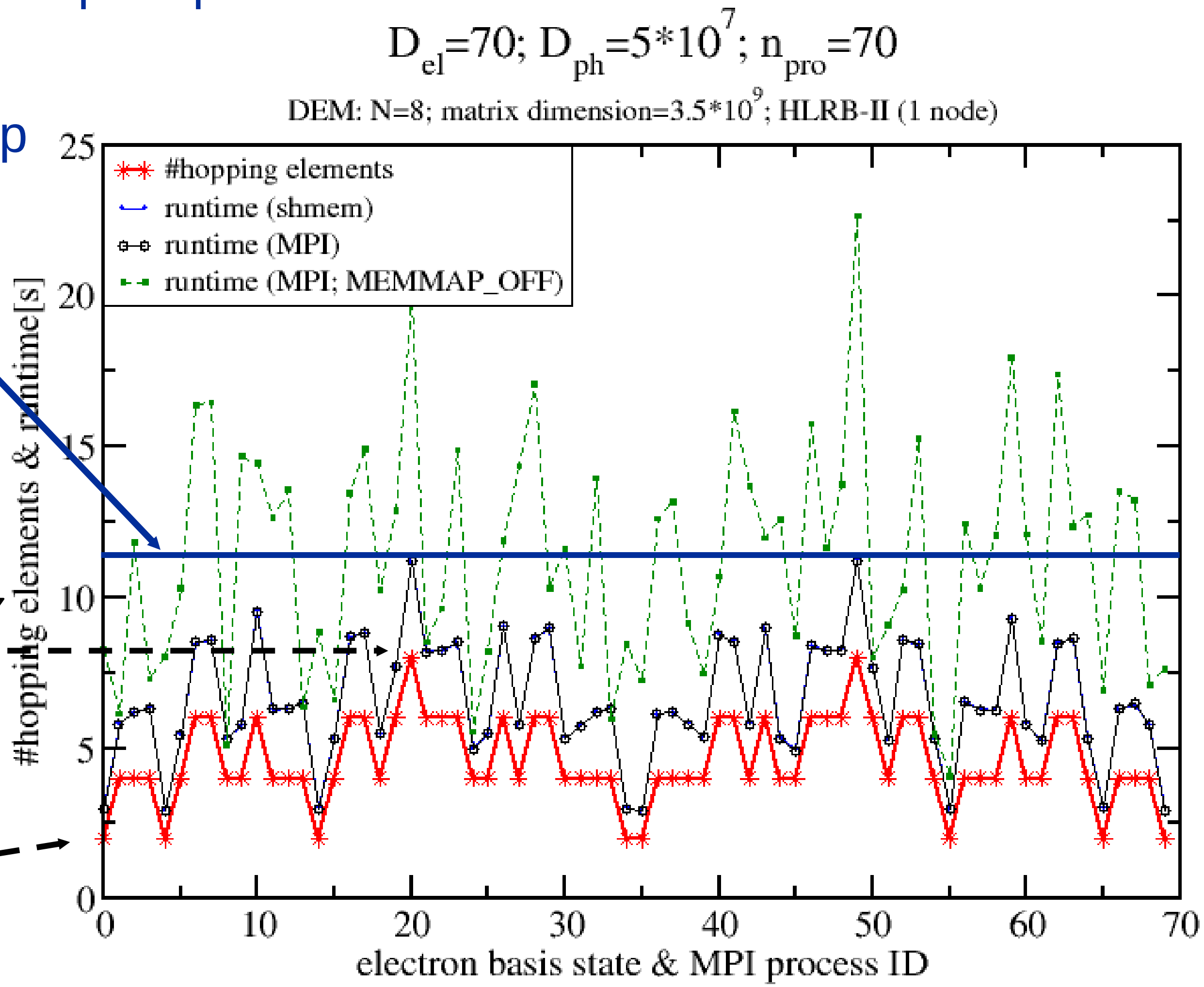


Improving Scalability:
Load imbalance

- Parallel performance is limited by load imbalances if $n_{pro} \sim D_{el}$
- Example: 8 sites; 4 spin-up electrons

Time per MVM step is determined by slowest process





Improving Scalability
"Transpose" basis set

- Alternatively: Distribute "**phonon** part" to n_{pro} processes

$\{v_{e,p};\ \mathbf{p}=(\mathbf{rank}*(D_{ph}/n_{pro})+1,\ldots,(\mathbf{rank}+1)*(D_{ph}/n_{pro});\ \mathbf{e=1,\ldots,D_{el}}\}$

-> Severe load imbalances for phonon part

-> Process local operation for electron part

$$H = -t \sum_{\langle ij\rangle,\sigma} \left[c^\dagger_{i\sigma} c_{j\sigma} + \mathrm{H.c.}\right] + U \sum_i n_{i\uparrow} n_{i\downarrow} + g\omega_0 \sum_{i,\sigma} (b^\dagger_i + b_i) n_{i\sigma} + \omega_0 \sum_i b^\dagger_i b_i$$

Compute: Electron part

Transpose Basis

Compute: Phonon part

Transpose Basis

- Distribute "**electronic** part" to n_{pro} processes

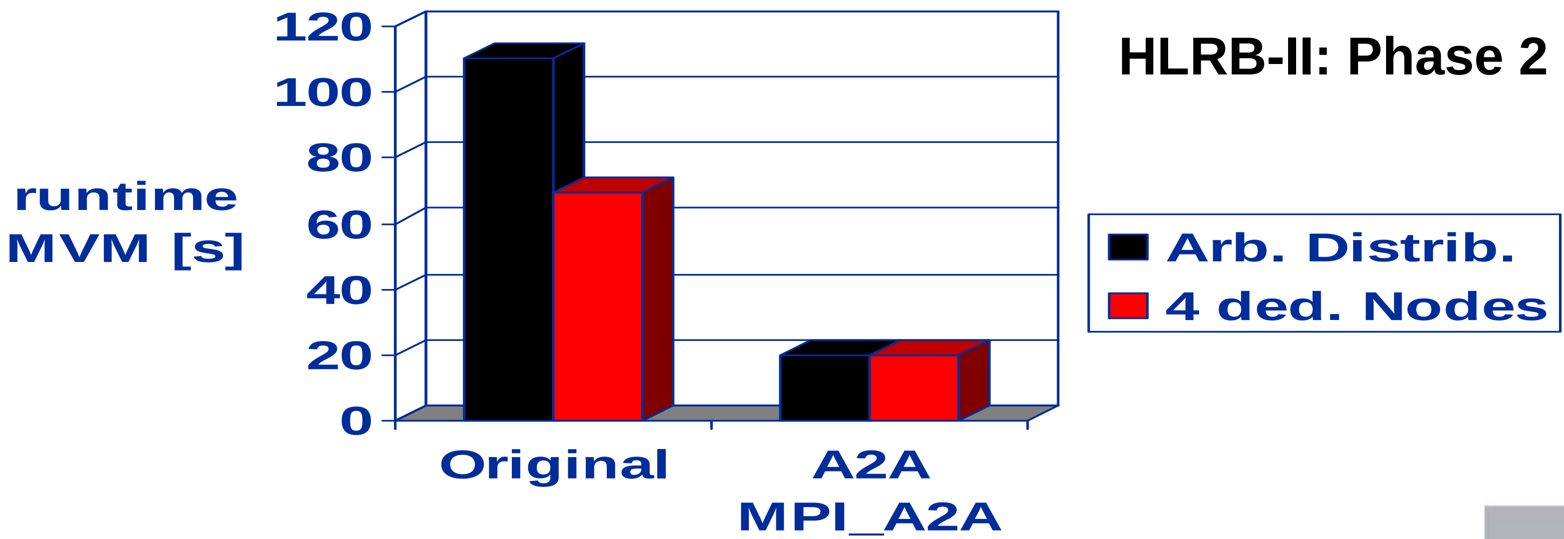
$\{v_{e,p};\ \mathbf{e}=(\mathbf{rank}*(D_{el}/n_{pro})+1,\ldots,(\mathbf{rank}+1)*(D_{el}/n_{pro});\ \mathbf{p=1,\ldots,D_{ph}}\}$

-> Severe load imbalances for electron part

-> Process local operation for phonon part

Improving Scalability
Transpose basis set using MPI_AllToAll

- Communication requirements: 2 MPI_Alltoall (A2A) per MVM
- A2A implementation improves load balancing and reduces maximum data transfer per MPI process, e.g. for $D_{el}/n_{pro}=1$
  - A2A implementation: $2 * 2 *$ $(n_{pro}-1)/n_{pro} * D_{ph} * 8$ Byte
  - Original implementation: $2 * 2 * dim * N_{el}$ $* D_{ph} * 8$ Byte
    ($dim=1,2$; $N_{el}=1,\ldots,N/2$ with $N=8,\ldots,16$)
- Test case: $N=16$, $N_{el}=4$, $dim=2$; $D_{el}=1820$; $D_{ph}=30*10^6$; 1820 cores

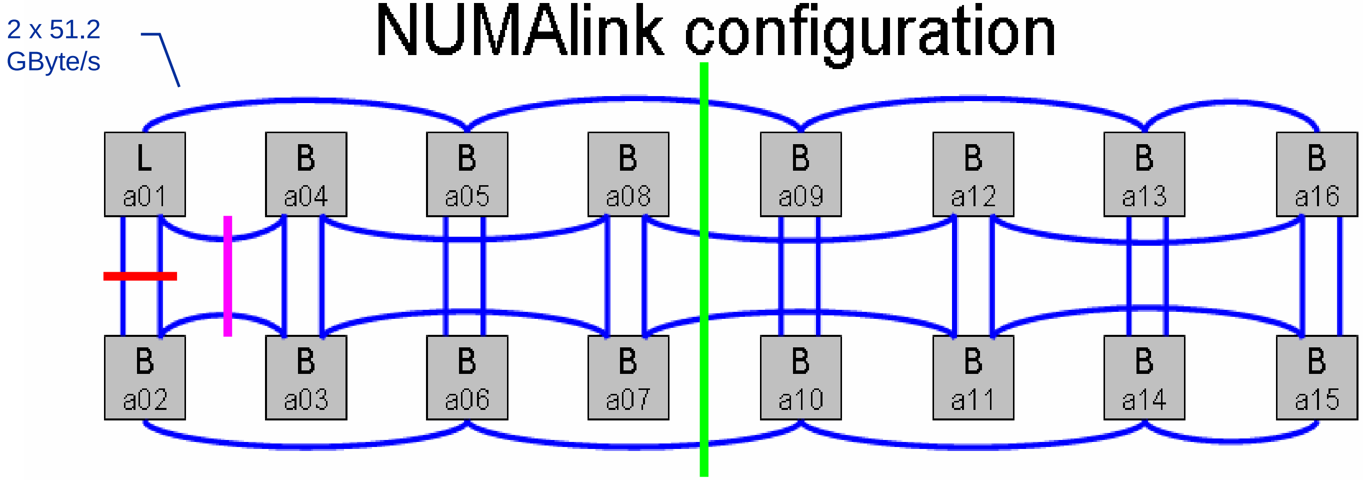


Improving Scalability

MPI_Alltoall and bisectional bandwidth of HLRB-II

- Test case: MPI_alltoall ~ 8.8 s at 240 MB vector -> 27 MB/s per core
- Minimum available bandwidth > 50 MB/s per core & direction

Inter-Partition NUMAlink configuration



Each grey box:
- 256 socket partition (SSI)
- L login, B batch

Each line represents:
- 2 NUMAlink4 planes with 16 cables (total)
- each cable: 2 * 3.2 GB/s

Bisection Bandwidths per socket pair:
- intra-partition: 2 * 0.8 GB/s
- any 2 vertical partitions: 2 * 0.4 GB/s
- 4 partitions: 2 * 0.2 GB/s
  many hops for bad choice → less bandwidth
- total system: **2 * 0.1 GB/s**

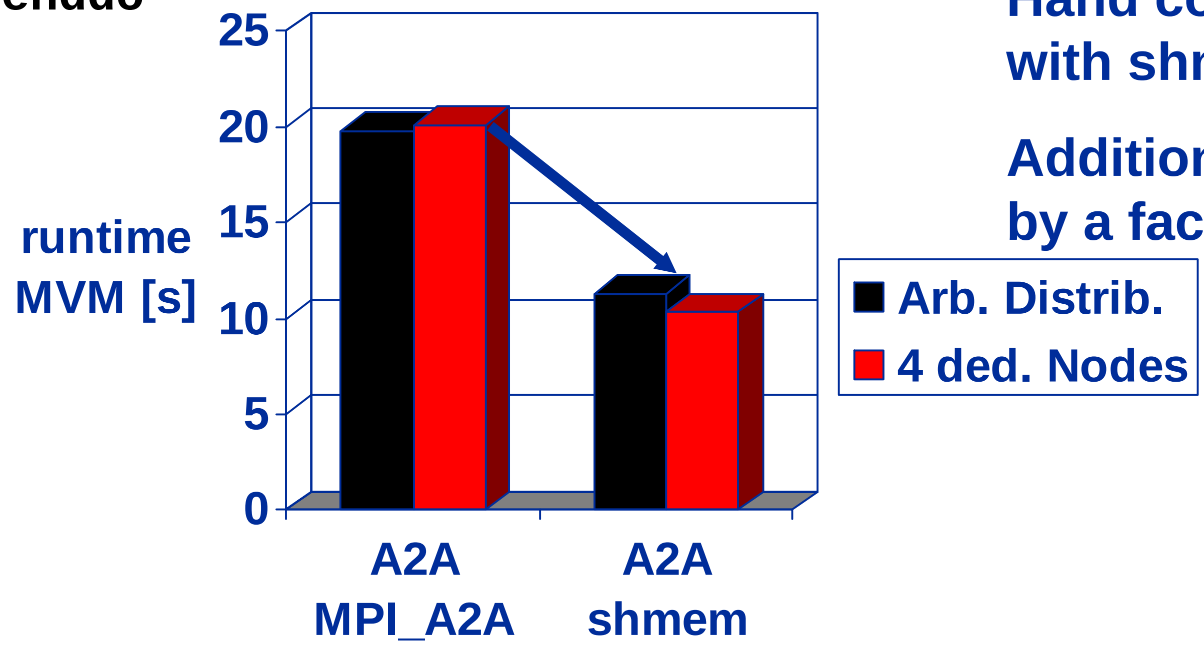
(factor 2 indicates "per direction")

by courtesy of LRZ

Improving Scalability
Improving A2A communication: shmem_get

- **Replace *MPI_Alltoall* call by explicit *shmem_get* calls**
- **Shift *shmem_get* calls to avoid network contention**

```
do r_part = 1 , (npro-1)
   get_data_from = mod(r_part + RANK, npro)
   call shmem_get( local , remote , … , get_data_from)
enddo
```

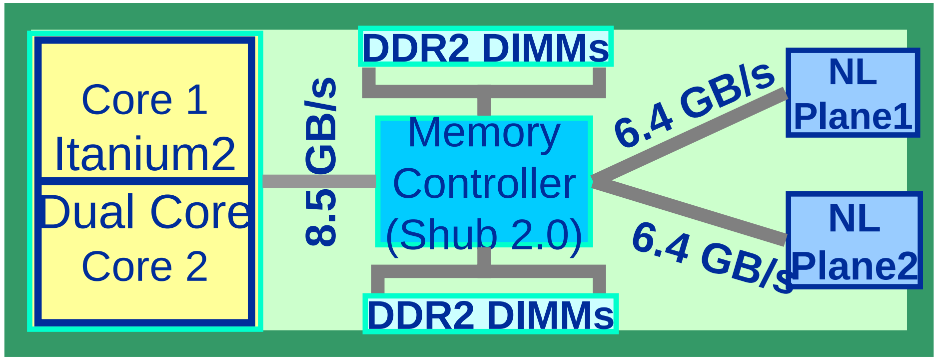


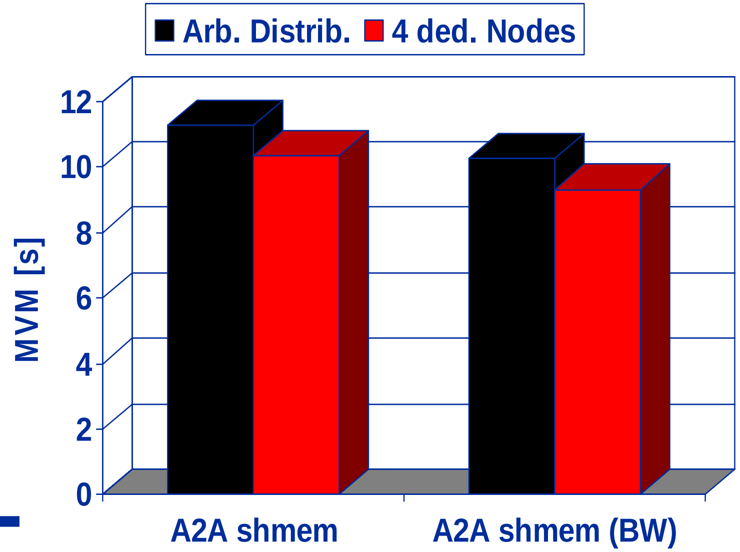
Hand coded A2A with shmem:

Additional speed up by a factor of 2 !

Improving Scalability

Black & White strategy to reduce network contention





Even ranks	*Odd ranks*
Transpose Basis	*Compute:*
Compute:	*Phonon part*
Electron part	**Transpose Basis**
Transpose Basis	*Compute:*
Compute:	*Electron part*
Phonon part	**Transpose Basis**

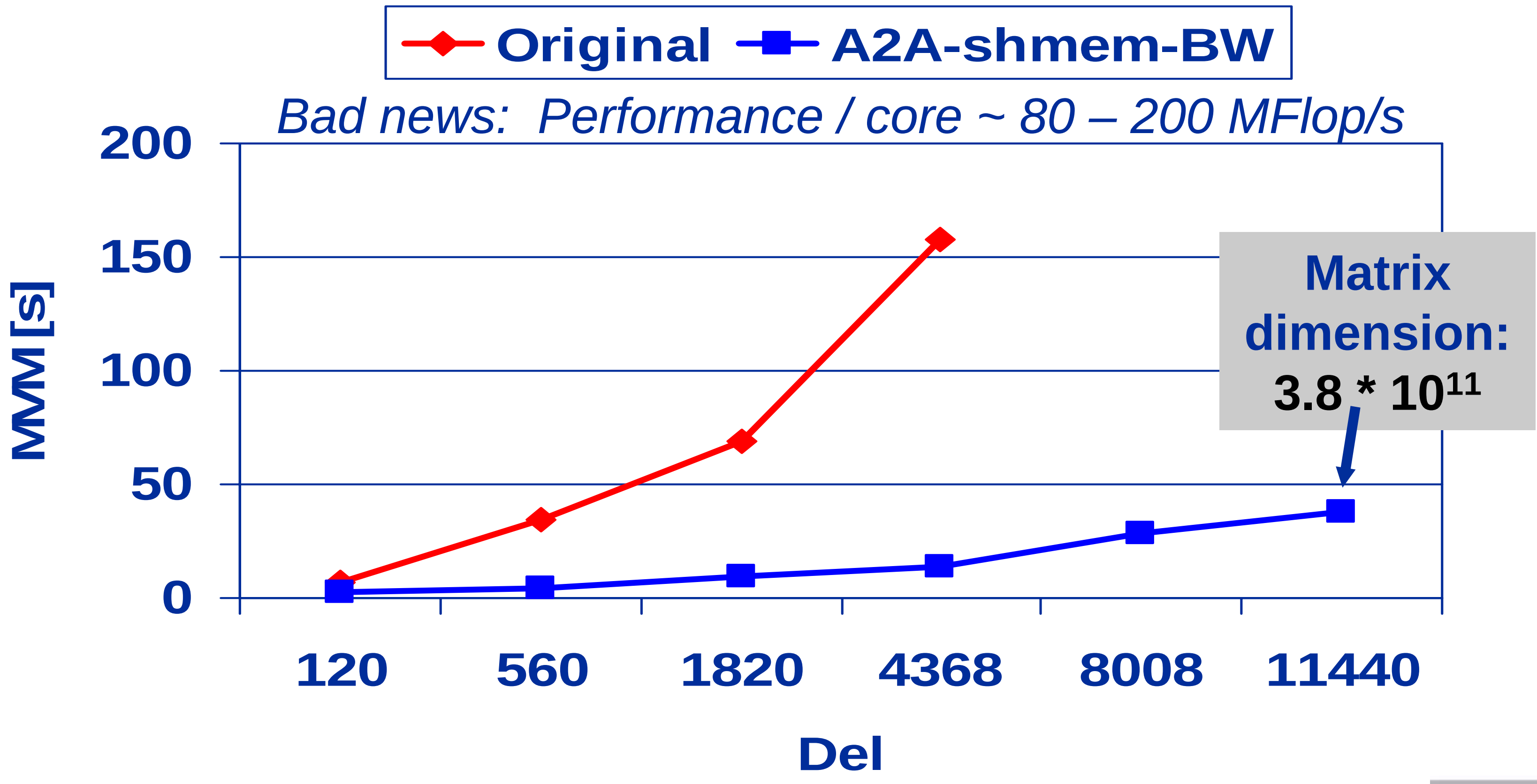
Gain is proportional to compute time for phonon part!

Test case:
Compute phonon ~1 s

Improving Scalability

Towards very large scale ED studies

- "Weak scaling" analysis with different number of electrons, i.e. D_{el}
- $D_{el}/n_{pro}=1$ (D_{el}=120,560,1820,4368) - $D_{el}/n_{pro}=2$ (D_{el}=8008,11440)



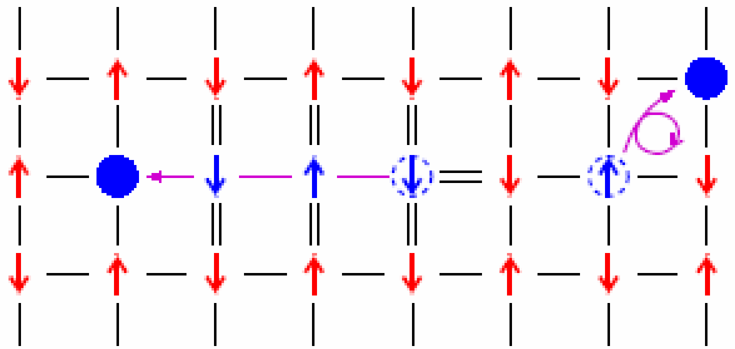
Current progress in physics

New model for boson assisted hopping transport
(D. Edwards, Imperial College)

$$H = -t_b \sum_{\langle i,j \rangle} c_j^\dagger c_i (b_i^\dagger + b_j) - \lambda \sum_i (b_i^\dagger + b_i) + \omega_0 \sum_i b_i^\dagger b_i + \frac{N\lambda^2}{\omega_0}$$

hopping   boson relaxation   boson energy

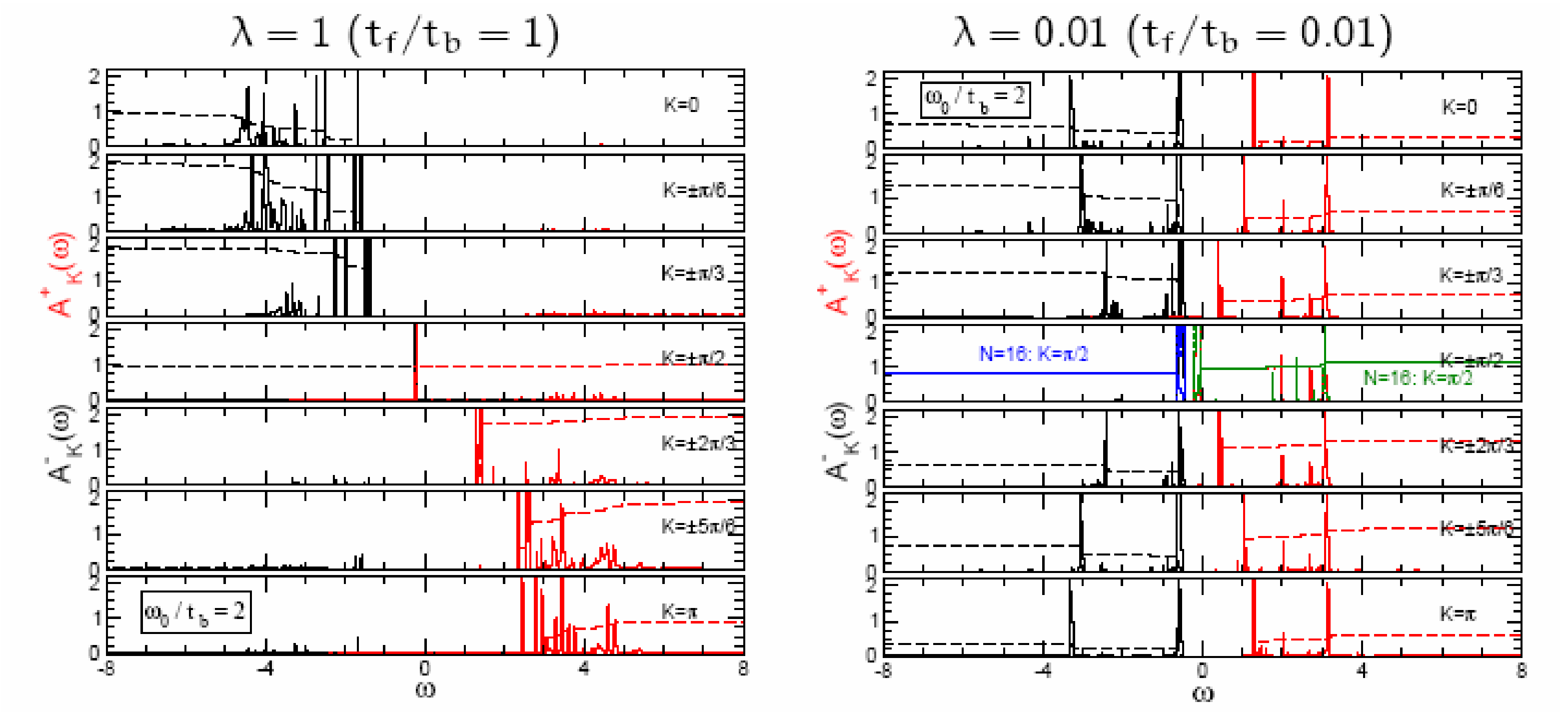
High-T_c cuprates (AFM spin background)



classical spins: "string effect"
hole is bound to its starting point
quantum spins: "fluctuations"
spin lattice can heal itself with rate
controlled by exchange parameter
$\rightsquigarrow$ $t - J-$ type models

Current progress in physics

- Variational ED approach with 1 & 2 particles:
  A. Alvermann, D.M. Edwards, H. Fehske, Phys. Rev. Lett. 88, 056602 ('07)
- Spectral properties at half-filling (ED studies - 500+ cores on HLRB-II)



• Metal – insulator transition as λ decreases

Thanks to our collaborators

- A. R. Bishop (Los Alamos National Lab, USA)
- D. Edwards (Imperial College, UK)
- E. Jeckelmann (Univ. Hannover, D)
- J. Loos (Czech Academy of Science, CZ)
- M. Hohenadler (Cambridge, UK)
- A. Basermann (NEC, St. Augustin, D)
- ….

Thanks to

- KONWIHR for funding the numerical & technical work
- LRZ Munich for providing access to a very powerful machine