Università della Svizzera italiana | Faculty of Informatics | Institute of Computational Science ICS

HP2C High Performance and High Productivity Computing

Parallel Scale-Transfer in Multiscale MD-FE Coupling using Remote Memory Access

D. Krause (dorian.krause@usi.ch), R. Krause (rolf.krause@usi.ch)
Institute of Computational Science, Università della Svizzera Italiana
Lugano, Switzerland

1st Workshop on Distributed Multiscale Computing
December 5, 2011, Stockholm, Sweden

Università della Svizzera italiana
Faculty of Informatics
Institute of Computational Science ICS

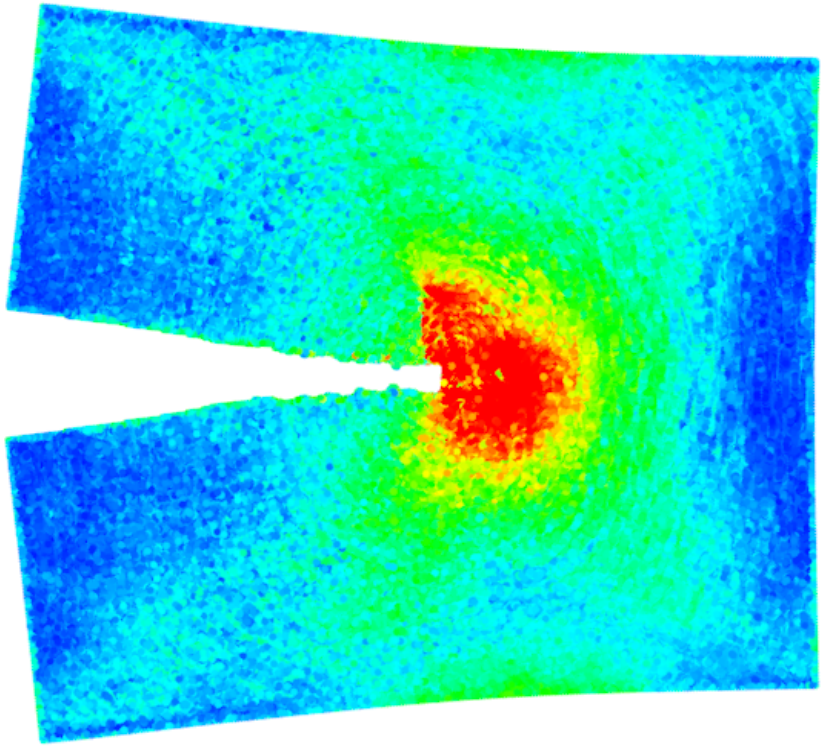
Multiscale Coupling

Concurrent Coupling

- Tackle cascade of scales with different models (Adaptivity) ✔
- Fine- and coarse model concurrently running in one simulation

Example: Fracture Mechanics

- Complicated physics near crack tip (hard to coarse grain)
- Spurious finite size effects ✗
- Molecular Dynamics good for near-crack region
- Elasticity good (enough) for rest ✔



Università della Svizzera italiana
Faculty of Informatics
Institute of Computational Science ICS

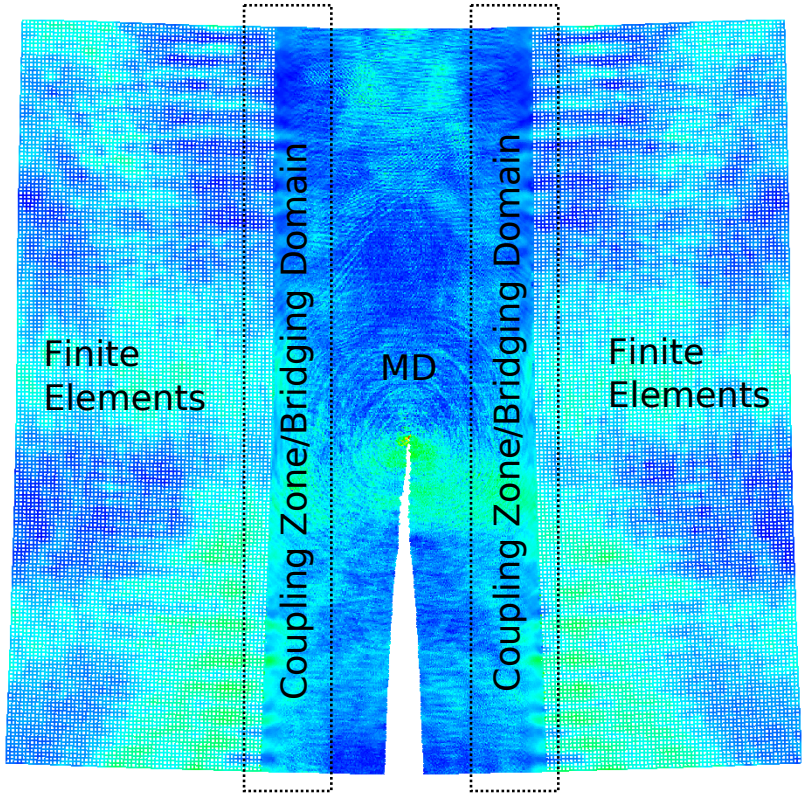
Weak Bridging Domain Method

Ingredients

- Overlapping domain decomposition
- Equations of motion derived from weighted Hamiltonian

$$\mathcal{H} = \alpha h + (1 - \alpha) H + \boldsymbol{\lambda} \cdot \mathbf{G}$$

- Constraints $\mathbf{G} = \mathbf{U} - P\mathbf{u}$
- P projection operator (Least squares or L^2 projection)
- RATTLE time integration requires solution of two linear systems in each time step ✗



- Damping of high fluctuation field using Perfectly Matched Layer method

Weak Bridging Domain Method

1 Compute trial values $\mathbf{u}^*$, $\mathbf{v}^*$, $\mathbf{U}^*$, $\mathbf{V}^*$ by applying the usual "Verlet kick" and "Verlet drift" ignoring the Lagrange forces:

$$\begin{bmatrix}\mathbf{v}^* \\ \mathbf{V}^*\end{bmatrix} = \begin{bmatrix}\mathbf{v}^n \\ \mathbf{V}^n\end{bmatrix} + \frac{\tau}{2}\begin{bmatrix}\mathbf{m}^{-1}\mathbf{f}^{n+1} \\ \mathbf{M}^{-1}\mathbf{F}^{n+1}\end{bmatrix},$$

$$\begin{bmatrix}\mathbf{u}^* \\ \mathbf{U}^*\end{bmatrix} = \begin{bmatrix}\mathbf{u}^n \\ \mathbf{U}^n\end{bmatrix} + \tau\begin{bmatrix}\mathbf{v}^* \\ \mathbf{V}^*\end{bmatrix}$$

where $\mathbf{f}^n$, $\mathbf{F}^n$ denote the forces computed in step 4 of the previous time step.

2 Compute $\mathbf{G}^* = \mathbf{R}\mathbf{u}^* - \tilde{\mathbf{M}}\mathbf{U}^*$ and solve $\mathbf{G}^* = \Lambda\boldsymbol{\lambda}$ for $\boldsymbol{\lambda}$.

3 Computed corrected values

$$\begin{bmatrix}\mathbf{v}^{n+\frac{1}{2}} \\ \mathbf{V}^{n+\frac{1}{2}}\end{bmatrix} = \begin{bmatrix}\mathbf{v}^* \\ \mathbf{V}^*\end{bmatrix} + \frac{1}{\tau}\begin{bmatrix}\mathbf{m}^{-1}\mathbf{R}^\mathsf{T}\boldsymbol{\lambda} \\ -\mathbf{M}^{-1}\tilde{\mathbf{M}}^\mathsf{T}\boldsymbol{\lambda}\end{bmatrix}$$

$$\begin{bmatrix}\mathbf{u}^{n+1} \\ \mathbf{U}^{n+1}\end{bmatrix} = \begin{bmatrix}\mathbf{u}^* \\ \mathbf{U}^*\end{bmatrix} + \begin{bmatrix}\mathbf{m}^{-1}\mathbf{R}^\mathsf{T}\boldsymbol{\lambda} \\ -\mathbf{M}^{-1}\tilde{\mathbf{M}}^\mathsf{T}\boldsymbol{\lambda}\end{bmatrix}.$$

4 Evaluate forces $\mathbf{f}^{n+1}$, $\mathbf{F}^{n+1}$ according to the Hamiltonian equation (without constraints). The MD force $\mathbf{f}^{n+1}$ also contains a (linear) damping term and can be written as

$$\mathbf{f}^{n+1} = \mathbf{f}^{\text{int},n+1} + 2\mathbf{m}\mathbf{D}\mathbf{Q}\mathbf{v}^{n+\frac{1}{2}}.$$

5 Compute trial velocity values

$$\begin{bmatrix}\mathbf{v}^* \\ \mathbf{V}^*\end{bmatrix} = \begin{bmatrix}\mathbf{v}^{n+\frac{1}{2}} \\ \mathbf{V}^{n+\frac{1}{2}}\end{bmatrix} + \frac{\tau}{2}\begin{bmatrix}\mathbf{m}^{-1}\mathbf{f}^{n+1} \\ \mathbf{M}^{-1}\mathbf{F}^{n+1}\end{bmatrix}.$$

6 Compute $\dot{\mathbf{G}}^* = \mathbf{R}\mathbf{v}^* - \tilde{\mathbf{M}}\mathbf{V}^*$ and solve $\dot{\mathbf{G}}^* = \Lambda\boldsymbol{\mu}$ for $\boldsymbol{\mu}$.

7 Correct the velocities

$$\begin{bmatrix}\mathbf{v}^{n+1} \\ \mathbf{V}^{n+1}\end{bmatrix} = \begin{bmatrix}\mathbf{v}^* \\ \mathbf{V}^*\end{bmatrix} + \begin{bmatrix}\mathbf{m}^{-1}\mathbf{R}^\mathsf{T}\boldsymbol{\mu} \\ -\mathbf{M}^{-1}\tilde{\mathbf{M}}^\mathsf{T}\boldsymbol{\mu}\end{bmatrix}.$$

Weak Bridging Domain Method

1 Compute trial values $\mathbf{u}^*$, $\mathbf{v}^*$, $\mathbf{U}^*$, $\mathbf{V}^*$ by applying the usual "Verlet kick" and "Verlet drift" ignoring the Lagrange forces:

$$\begin{bmatrix}\mathbf{v}^* \\ \mathbf{V}^*\end{bmatrix} = \begin{bmatrix}\mathbf{v}^n \\ \mathbf{V}^n\end{bmatrix} + \frac{\tau}{2}\begin{bmatrix}\mathbf{m}^{-1}\mathbf{f}^{n+1} \\ \mathbf{M}^{-1}\mathbf{F}^{n+1}\end{bmatrix},$$

$$\begin{bmatrix}\mathbf{u}^* \\ \mathbf{U}^*\end{bmatrix} = \begin{bmatrix}\mathbf{u}^n \\ \mathbf{U}^n\end{bmatrix} + \tau\begin{bmatrix}\mathbf{v}^* \\ \mathbf{V}^*\end{bmatrix}$$

where $\mathbf{f}$ [illegible] in step 4 of the previous time step.

How to efficiently parallelize the scale-transfer?

2 Compute $\mathbf{G}^* = \mathbf{R}\mathbf{u}^* - \tilde{\mathbf{M}}\mathbf{U}^*$ and solve $\mathbf{G}^* = \Lambda\boldsymbol{\lambda}$ for $\boldsymbol{\lambda}$.

3 Computed corrected values

$$\begin{bmatrix}\mathbf{v}^{n+\frac{1}{2}} \\ \mathbf{V}^{n+\frac{1}{2}}\end{bmatrix} = \begin{bmatrix}\mathbf{v}^* \\ \mathbf{V}^*\end{bmatrix} + \frac{1}{\tau}\begin{bmatrix}\mathbf{m}^{-1}\mathbf{R}^\mathsf{T}\boldsymbol{\lambda} \\ -\mathbf{M}^{-1}\tilde{\mathbf{M}}^\mathsf{T}\boldsymbol{\lambda}\end{bmatrix}$$

$$\begin{bmatrix}\mathbf{u}^{n+1} \\ \mathbf{U}^{n+1}\end{bmatrix} = \begin{bmatrix}\mathbf{u}^* \\ \mathbf{U}^*\end{bmatrix} + \begin{bmatrix}\mathbf{m}^{-1}\mathbf{R}^\mathsf{T}\boldsymbol{\lambda} \\ -\mathbf{M}^{-1}\tilde{\mathbf{M}}^\mathsf{T}\boldsymbol{\lambda}\end{bmatrix}.$$

4 Evaluate forces $\mathbf{f}^{n+1}$, $\mathbf{F}^{n+1}$ according to the Hamiltonian equation (without constraints). The MD force $\mathbf{f}^{n+1}$ also contains a (linear) damping term and can be written as

$$\mathbf{f}^{n+1} = \mathbf{f}^{\text{int},n+1} + 2\mathbf{m}\mathbf{D}\mathbf{Q}\mathbf{v}^{n+\frac{1}{2}}.$$

5 Compute trial velocity values

$$\begin{bmatrix}\mathbf{v}^{n+1} \\ \mathbf{V}^{n+\frac{1}{2}}\end{bmatrix} = \begin{bmatrix}\mathbf{v}^{n} \\ \mathbf{V}^{n+\frac{1}{2}}\end{bmatrix} + \frac{\tau}{2}\begin{bmatrix}\mathbf{m}^{-1}\mathbf{f}^{n+1} \\ \mathbf{M}^{-1}\mathbf{F}^{n+1}\end{bmatrix}.$$

6 Compute $\dot{\mathbf{G}}^* = \mathbf{R}\mathbf{v}^* - \tilde{\mathbf{M}}\mathbf{V}^*$ and solve $\dot{\mathbf{G}}^* = \Lambda\boldsymbol{\mu}$ for $\boldsymbol{\mu}$.

7 Correct the velocities

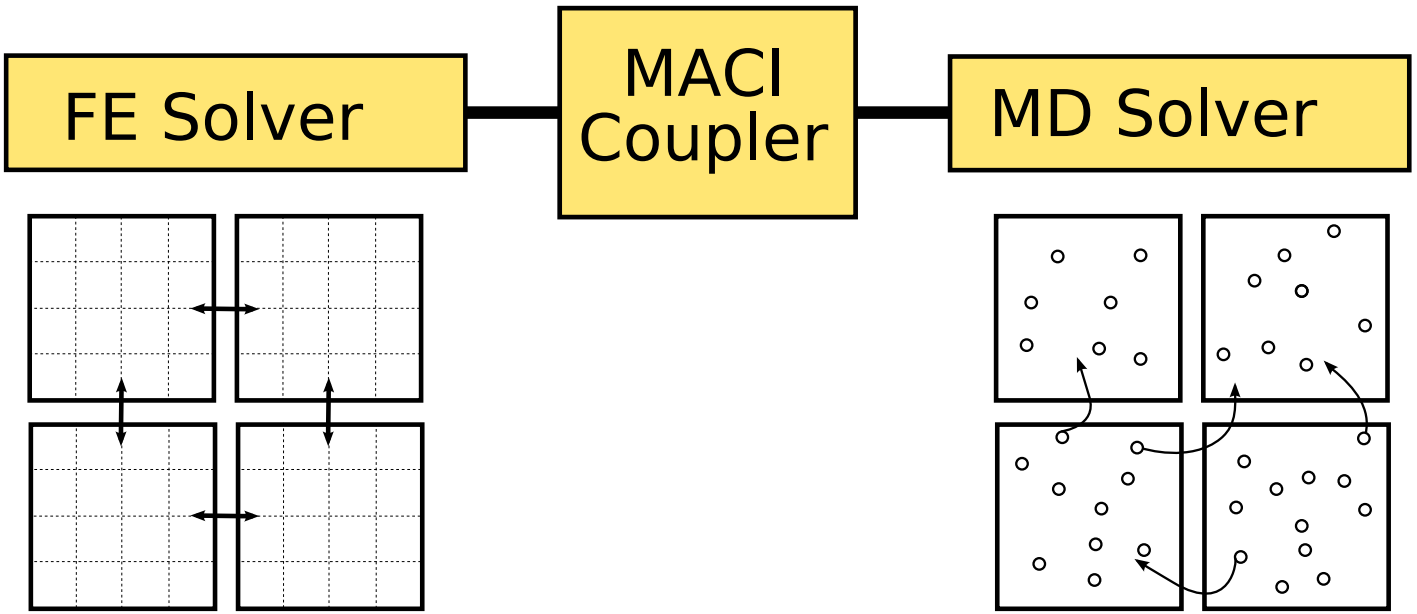
$$\begin{bmatrix}\mathbf{v}^{n+1} \\ \mathbf{V}^{n+1}\end{bmatrix} = \begin{bmatrix}\mathbf{v}^* \\ \mathbf{V}^*\end{bmatrix} + \begin{bmatrix}\mathbf{m}^{-1}\mathbf{R}^\mathsf{T}\boldsymbol{\mu} \\ -\mathbf{M}^{-1}\tilde{\mathbf{M}}^\mathsf{T}\boldsymbol{\mu}\end{bmatrix}.$$

Università della Svizzera italiana
Faculty of Informatics
Institute of Computational Science ICS

Software

MACI (Multiscale Atomistic Coupling Interface)

- Tool for the coupling of **commodity** MD and FE codes ✔
- Standardized interfaces for components
- MPMD-style execution (but single executable file)
- Data distribution dictated by components ✗



Università della Svizzera italiana
Faculty of Informatics
Institute of Computational Science ICS

Anatomy of Parallel Scale-Transfer

Scale-Transfer

Given particle fields $(\mathbf{u}_i)_i$, $(\mathbf{v}_i)_i$ compute

- $\mathbf{Z}_A = \sum_i \mathbf{R}_{Ai}\mathbf{u}_i$ MD ➜ FE
- $\mathbf{z}_i = \sum_A \mathbf{R}_{iA}\left[\Lambda^{-1}\left(\tilde{\mathbf{M}}\mathbf{U} - \mathbf{R}\mathbf{u}\right)\right]_A$ MD ➜ FE ➜ MD
- $\mathbf{z}_i = \sum_j \mathbf{Q}_{ij}\mathbf{v}_j$ MD ➜ MD ➜ MD

- Matrices distributed by rows
- MPMD execution prohibits data sharing ✗

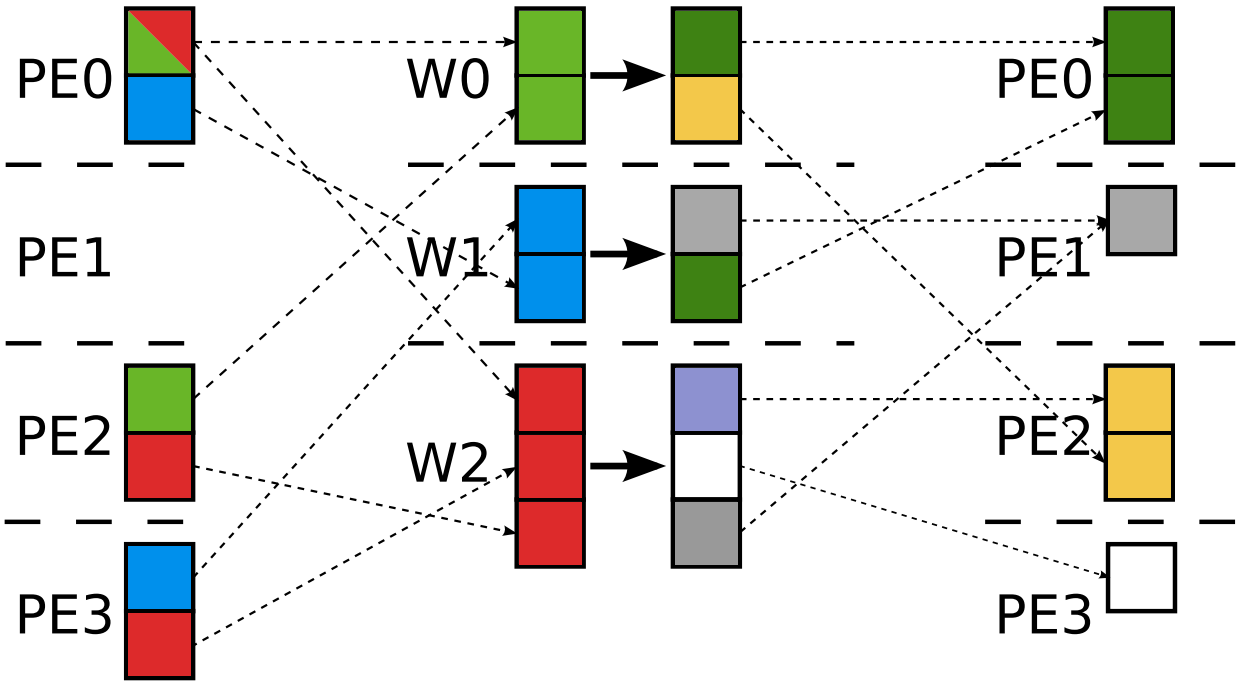
Challenges

- Dynamic data distribution (unknown a-priori) ✗
- Non-local nature of constraints ✗

Anatomy of Parallel Scale-Transfer

General Setup:

- Black-box operation $\mathrm{Op}_w : \mathbf{u} \to \mathbf{z}$ executed on set of **workers** $w \in G$
- I_w = tuple of indices such that the input to Op_w equals $\mathbf{u}(I_w) = (\mathbf{u}_i)_{i \in I_w}$
- O_w = tuple of indices such that the contribution of w to $\mathbf{z}$ equals $\mathbf{z}(O_w)$ (multiple contributions being summed up)



Anatomy of Parallel Scale-Transfer

Existing Approaches

- 1-dimensional decomposition ✗
- Event-based notification [Anciaux et al. '06]
  - ♯ Particle migration triggers notification of workers
  - ♯ Worker assigns new index to particles
  - ♯ Consistent ordering of send buffers and data layout on workers ✔
  - ♯ Changing data layouts on workers ✗

Our Approach

- Data distribution transparent to worker (good for modularity) ✔
- Assume MD processes know target workers and offsets for local particles
  - ♯ input targets $= \left\{ w \in G \;\middle|\; i \in I_w \right\}$
  - ♯ local input index k such that $(I_w)_k = i$
  - ♯ Piggyback as metadata onto particles ✔
- Embrace/Deal with **one-sided** nature of problem

Implementation Options

Algorithms

- **2-sided:** Exchange of metadata precedes exchange of data
  - Requires exchange of offsets ✗
  - Packing/unpacking of data on sender and receiver (performance insensitive to data sorting) ✔
  - **MPI Alltoall:** **1.** Use collectives to exchange #values to be send/recv'ed
    2. Allocate buffer space and use `MPI_Alltoallv` or point-to-point communication for exchanging particles and offsets
  - **MPI Pt2Pt :** **1.** Exchange offest using point-to-point communication with receives probing (`MPI_Probe`) in loop
    2. Exchange data with non-blocking send/recv calls

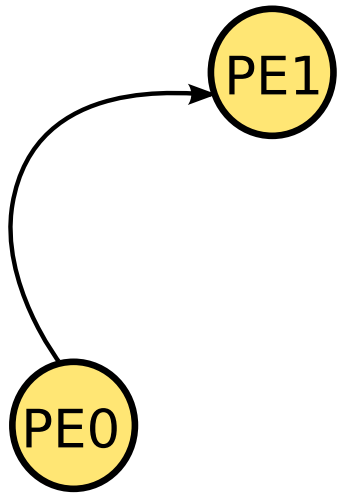
Implementation Options

Algorithms

- **2-sided:** Exchange of metadata precedes exchange of data
  - ♯ Requires exchange of offsets ✗
  - ♯ Packing/unpacking of data on sender and receiver (performance insensitive to data sorting) ✔
- **1-sided:** No explicit metadata exchange

Remote Memory Access

- Send and receive initiated by **origin**
- **Target** only implicitly involved (synchronization)
- Well suited for state-of-the-art interconnects with **RDMA** support
- Programming models: MPI-2 RMA, Global Arrays, (Open)SHMEM, ...



Implementation Options

Algorithms

- **2-sided:** Exchange of metadata precedes exchange of data
  - Requires exchange of offsets ✗
  - Packing/unpacking of data on sender and receiver (performance insensitive to data sorting) ✔
- **1-sided:** No explicit metadata exchange
  - No exchange of offsets ✔
  - Performance sensitive to data sorting ✗
  - Common structure: Origin processes put into or get data from **RMA exposed** memory + Collective synchronization
  - **MPI RMA:** Origin processes put or get data using `MPI_Put` and `MPI_Get` Collective `MPI_Fence` synchronization
  - **GA:** Similar to **MPI RMA** but with transparent distribution of the global array
  - **SHMEM:** Similar to **MPI RMA** but using `shemm_put`, `shmem_get` and `shmem_barrier_all`

Binning Benchmark

Binning

- Communication benchmark (weak scaling)
- Match particles to cells

Parameters

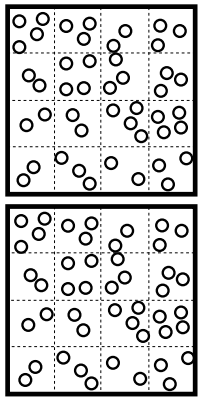
- N processes, W workers
- Number of particles per cell K
  - ♯ Controls number of put's/get's and message sizes
- Number of cells per worker per dimension L
  - ♯ Used to fix total number of particles

Configurations

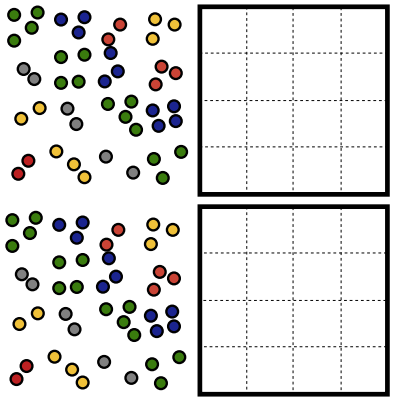
C1	$L = 64$	$K = 32$	✗
C2	$L = 32$	$K = 256$	
C3	$L = 16$	$K = 2048$	✔

- Number of particles per worker $\sim$ 8M

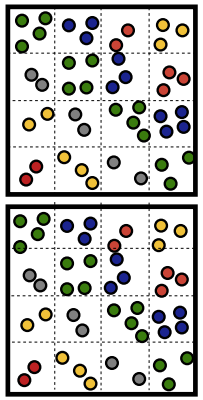
1



2



3

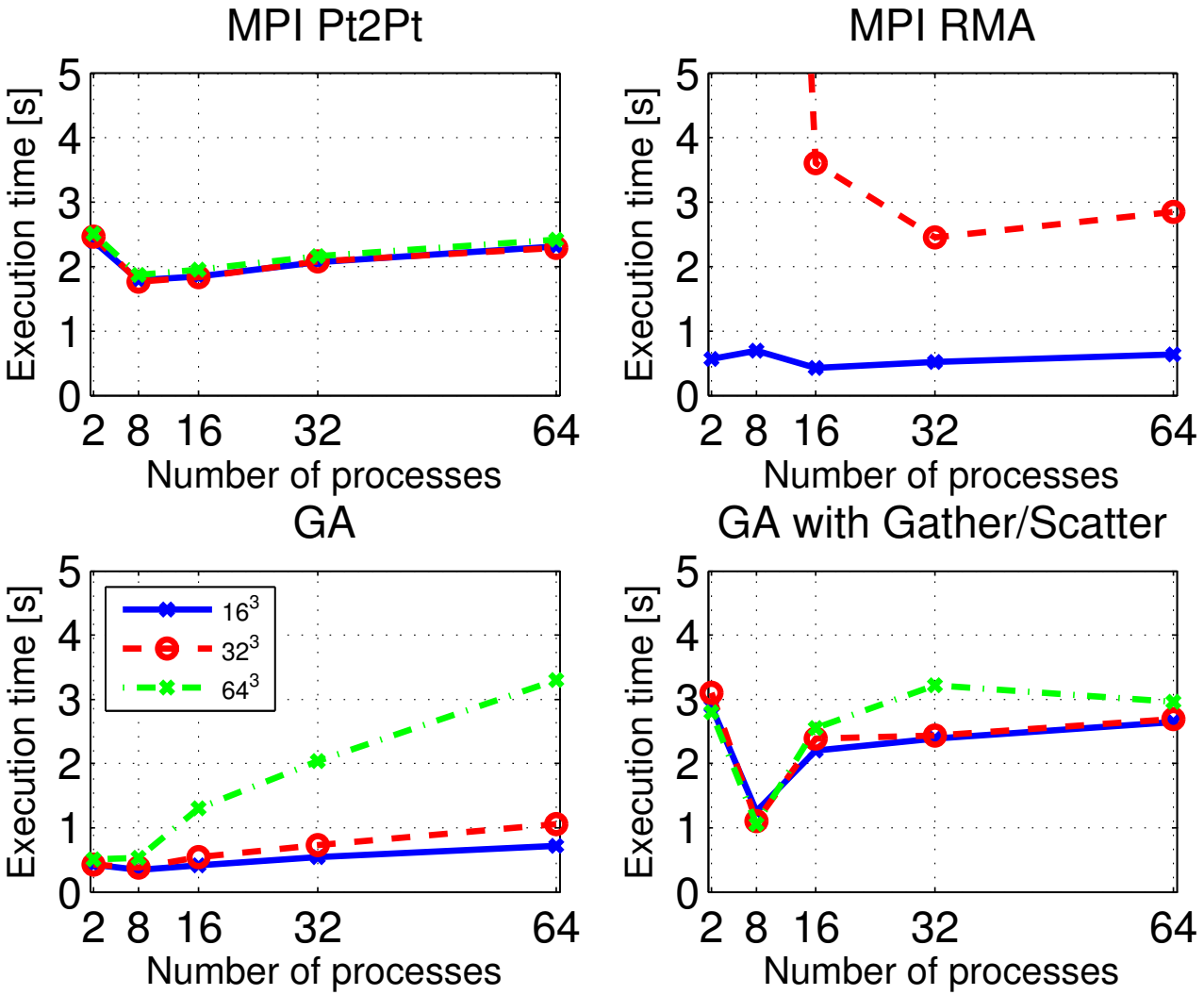


Test Systems

IB Cluster	Cray XT5	Cray XE6
Dual-socket quad-core AMD Opteron "Barcelona" nodes	Dual-socket hexa-core AMD Opteron "Istanbul" nodes	Dual-socket 12-core AMD Opteron "Istanbul" nodes
4x DDR Infiniband interconnect (RDMA support ✔)	**Seastar2+ interconnect** (optimized for MPI-1 subset ✗)	**Gemini interconnect** (RDMA support ✔)
Open MPI 1.4.2, Global Array 5.0.1	Cray MPT, Global Array 4.3.2	Cray MPT, Global Array 4.3.2
Every 8^{th} process as worker (one per node)	Every 12^{th} process as worker (one per node)	Every 12^{th} process as worker (two per node)

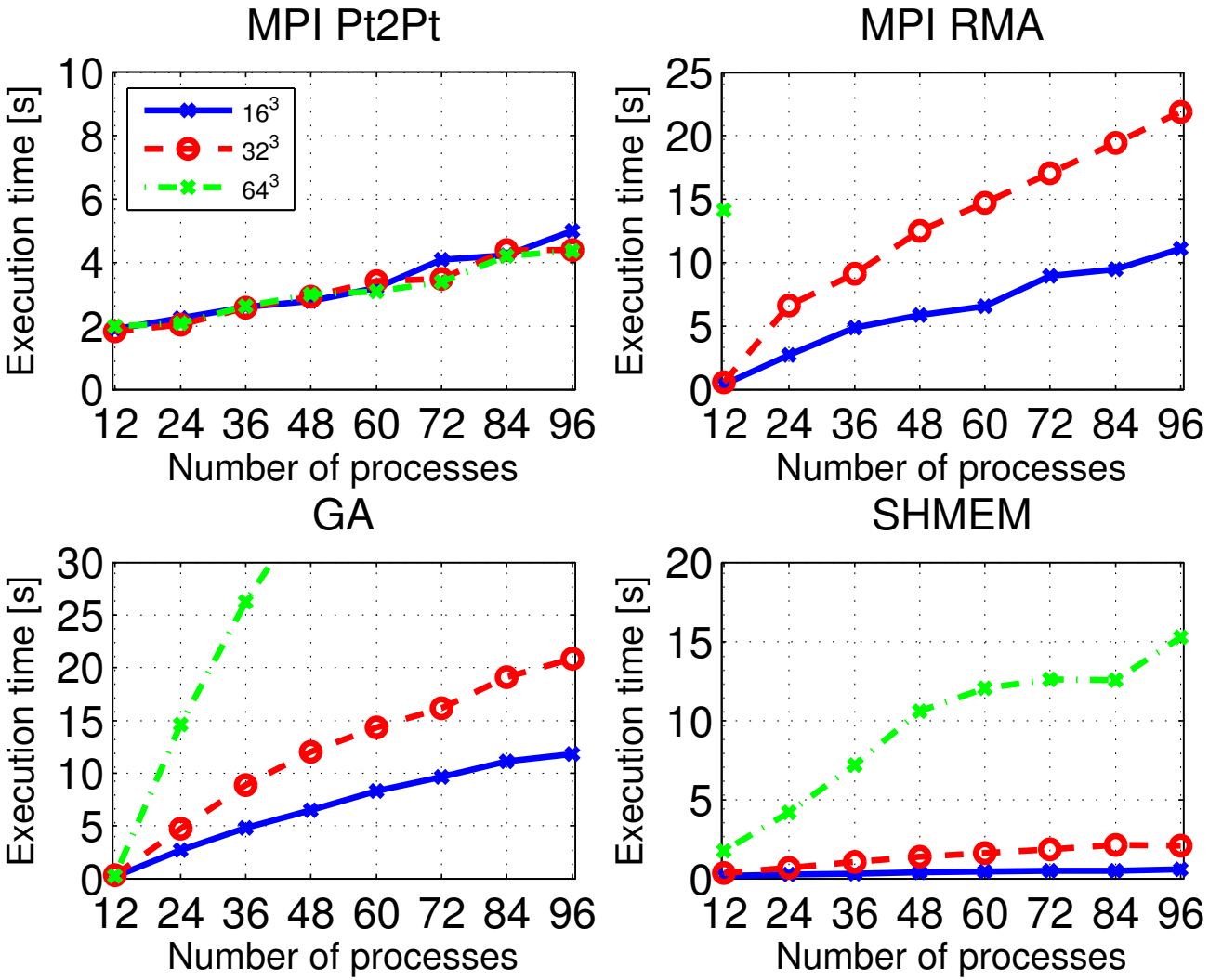
Università della Svizzera italiana
Faculty of Informatics
Institute of Computational Science ICS

Selected Results (IB Cluster)



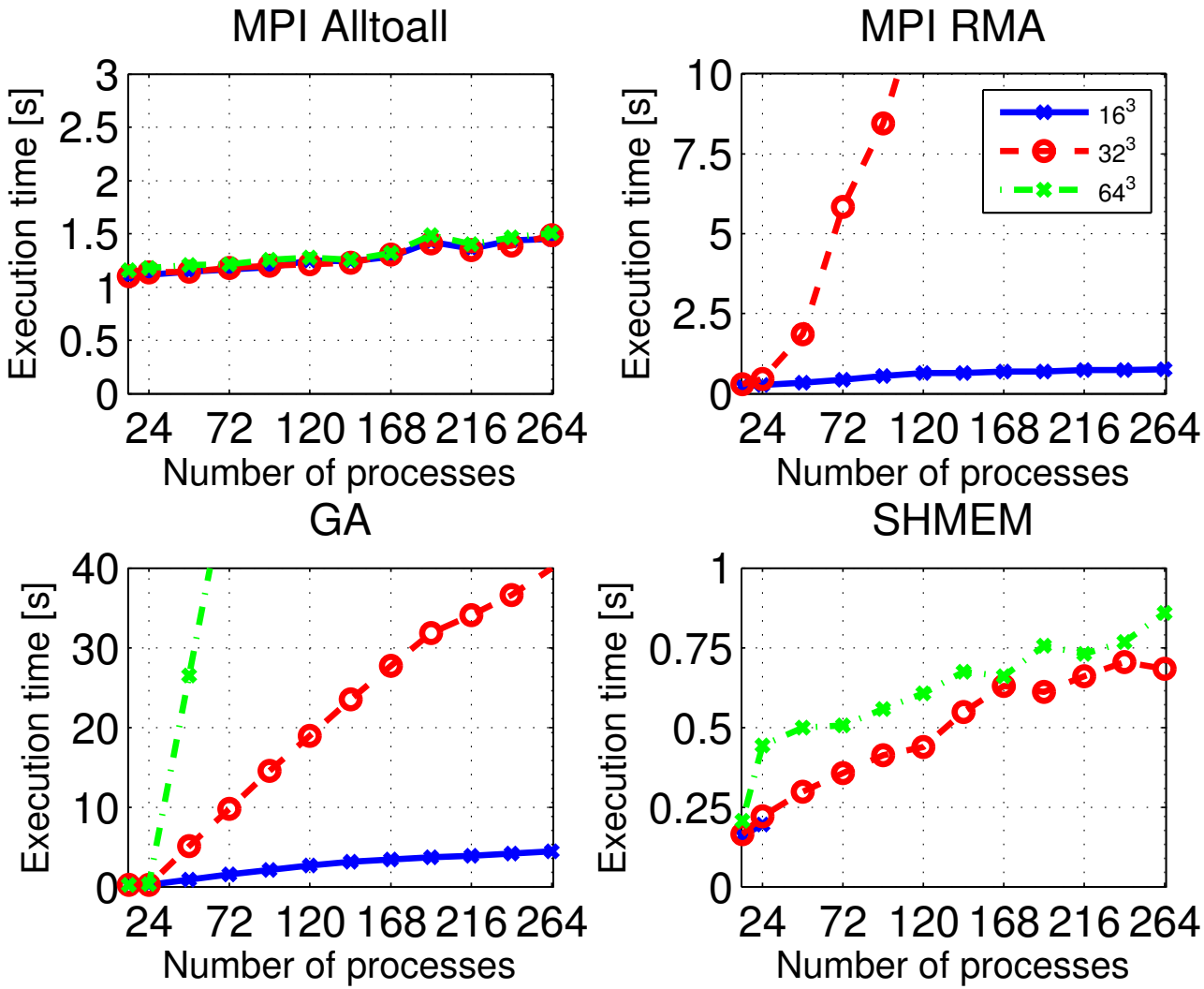
Università della Svizzera italiana
Faculty of Informatics
Institute of Computational Science ICS

Selected Results (Cray XT5)



Università della Svizzera italiana
Faculty of Informatics
Institute of Computational Science ICS

Selected Results (Cray XE6)



Università della Svizzera italiana | Faculty of Informatics | Institute of Computational Science ICS

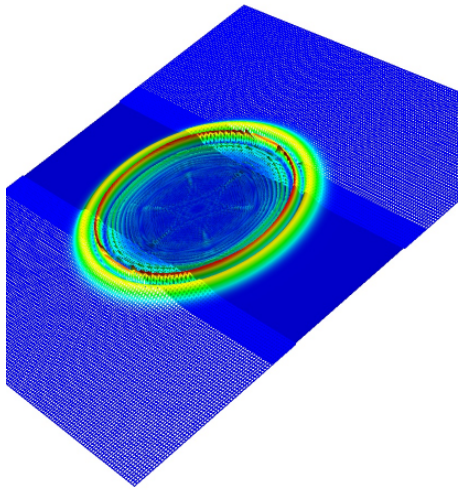
Conclusions

Multiscale MD-FE Coupling for Fracture Mechanics

- Molecular Dynamics for near crack tip region, continuum theory for remainder
- Model adaptive ✔
- Coupling with **averaging constraints**

Novel Parallelization Approach for Parallel Scale-Transfer

- Piggyback'ed metadata
- Algorithms respect data ordering on FE processes ✔
- Data distribution transparent to workers ✔
- Good match for RDMA capable interconnects ✔
- Performance depends on data order ✗



Software

- C/C++/Python tool for Linux clusters
- Tested components: UG FE code, Tremolo and LAMMPS MD codes

Strong Scaling MACI