GPU-Accelerated Particle-in-cell Code on Minsky

IWOPH17, ISC, Frankfurt a. M.

Andreas Herten, Forschungszentrum Jülich, 22 July 2017

Outline

Contributions *TL;DR*

- PiC Code to GPU (partly)
- OpenACC, CUDA Fortran
- Data layout benchmarks on Minsky (POWER8*NVL*, P100)
- Peculiarities with PGI compiler on POWER
- Performance Model

Jülich Supercomputing Centre

Part of Forschungszentrum Jülich

- Forschungszentrum Jülich
  - One of Europe's largest research centers (≈6000 employees)
  - Energy, environmental sciences, health, information technology
- Jülich Supercomputing Centre
  - Two Top 500 supercomputers (JUQUEEN: #21, JURECA: #80)
  - NVIDIA Application Lab
  - POWER Acceleration and Design Centre

*Super*computers Involved

JÜLICH FORSCHUNGSZENTRUM

JURON **JURECA** **JUHYDRA**

*Super*computers Involved

JURON

- Human Brain Project prototype
- 18 nodes with IBM POWER8*NVL* CPUs (2×10 cores)
- Per Node: 4 NVIDIA Tesla P100 cards, connected via NVLink.
- GPU: 0.38 PFLOP/s peak performance
- NVME

JURECA

JUHYDRA

*Super*computers Involved

JURON

- Human Brain Project prototype
- 18 nodes with IBM POWER8*NVL* CPUs (2×10 cores)
- Per Node: 4 NVIDIA Tesla P100 cards, connected via NVLink.
- GPU: 0.38 PFLOP/s peak performance
- NVME

JURECA

- General-purpose supercomputer
- 1872 nodes with Intel Xeon E5 CPUs (2×12 cores)
- 75 nodes with 2 NVIDIA Tesla K80 cards
- 1.8 (CPU) + 0.44 (GPU) PFLOP/s peak performance (#70)
- EDR InfiniBand

JUHYDRA

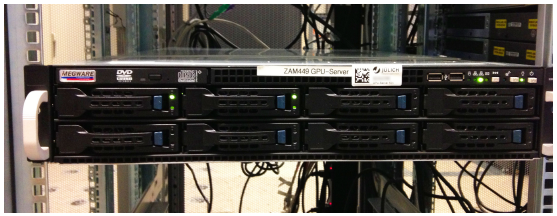
*Super*computers Involved

JURON

- Human Brain Project prototype
- 18 nodes with IBM POWER8*NVL* CPUs (2×10 cores)
- Per Node: 4 NVIDIA Tesla P100 cards, connected via NVLink.
- GPU: 0.38 PFLOP/s peak performance
- NVME

JURECA

- General-purpose supercomputer
- 1872 nodes with Intel Xeon E5 CPUs (2×12 cores)
- 75 nodes with 2 NVIDIA Tesla K80 cards
- 1.8 (CPU) + 0.44 (GPU) PFLOP/s peak performance (#70)
- EDR InfiniBand



JUHYDRA

- GPU prototyping machine
- 1 node with Intel Xeon E5 CPU (2×8 cores)
- NVIDIA $2 \times$ Tesla K20, $2 \times$ Tesla K40 cards
- No batch system

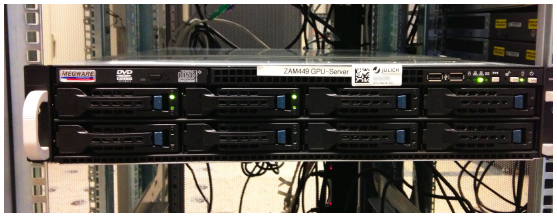
*Super*computers Involved

JURON

- Human Brain Project prototype
- 18 nodes with IBM POWER8*NVL* CPUs (2×10 cores)
- Per Node: 4 NVIDIA Tesla P100 cards, connected via NVLink.
- GPU: 0.38 PFLOP/s peak performance
- NVME

JURECA

- General-purpose supercomputer
- 1872 nodes with Intel Xeon E5 CPUs (2×12 cores)
- 75 nodes with 2 NVIDIA Tesla K80 cards
- 1.8 (CPU) + 0.44 (GPU) PFLOP/s peak performance (#70)
- EDR InfiniBand



JUHYDRA

- GPU prototyping machine
- 1 node with Intel Xeon E5 CPU (2×8 cores)
- NVIDIA $2 \times$ Tesla K20, $2 \times$ Tesla K40 cards
- No batch system

JuSPIC

JuSPIC

A scalable Particle-in-Cell plasma physics code

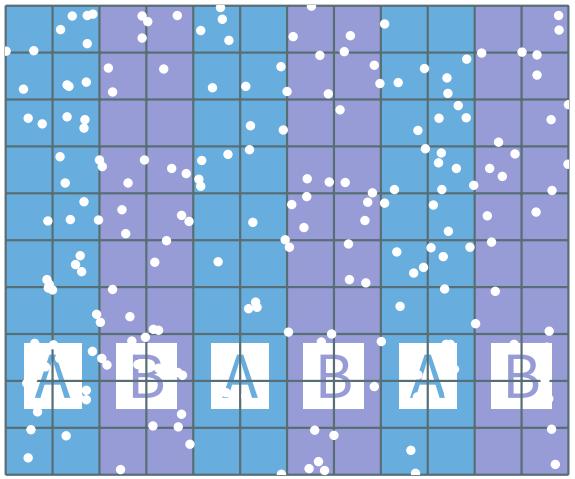
- Based on PSC by H. Ruhl
- Laser-plasma interaction
- 3D electromagnetic PiC code
- Finite-Difference Time-Domain scheme
- Cartesian geometry, arbitrary number of particle species
- Scales to full Blue Gene/Q system JUQUEEN

JuSPIC

A scalable Particle-in-Cell plasma physics code

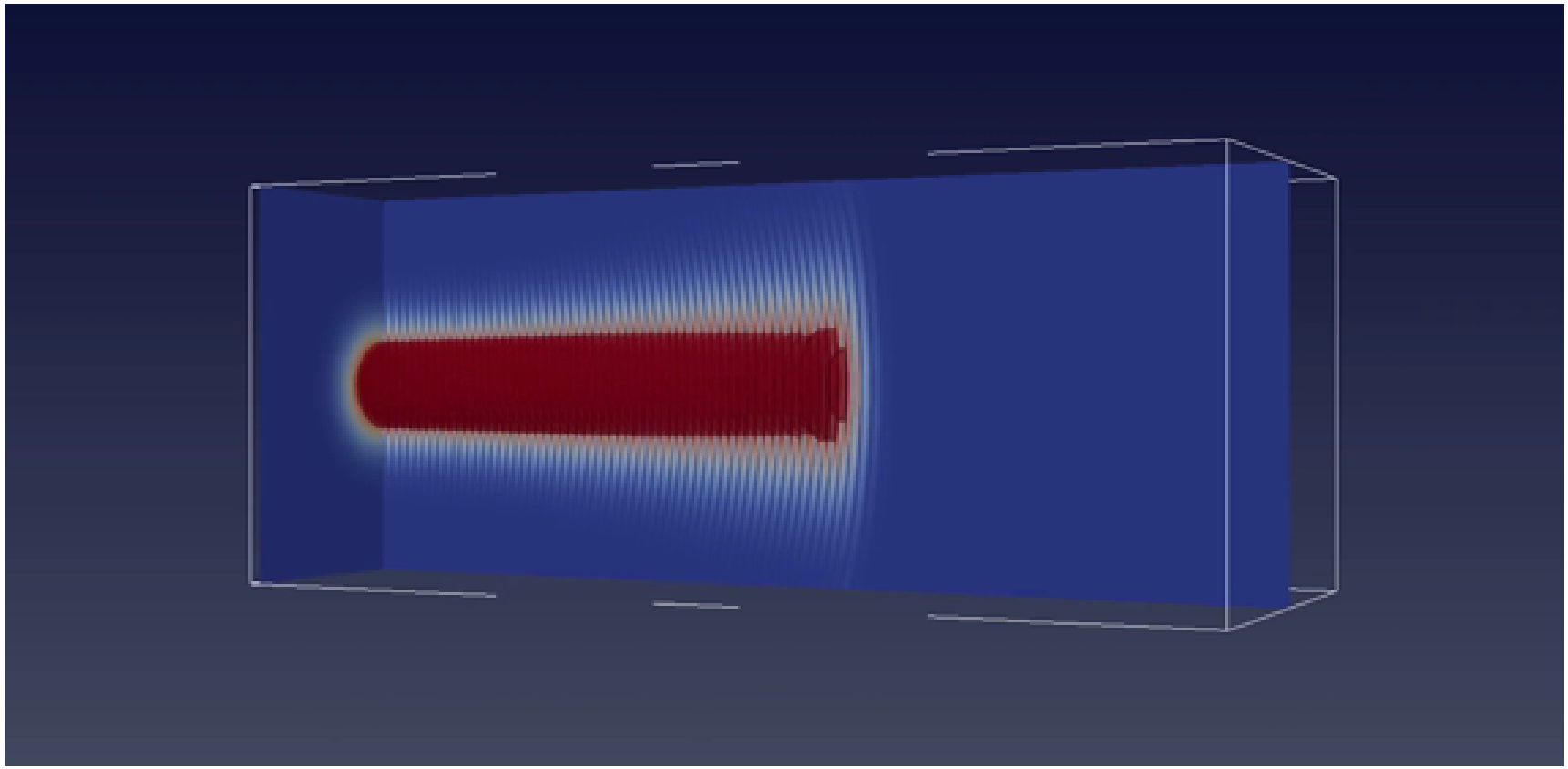
- Based on PSC by H. Ruhl
- Laser-plasma interaction
- 3D electromagnetic PiC code
- Finite-Difference Time-Domain scheme
- Cartesian geometry, arbitrary number of particle species
- Scales to full Blue Gene/Q system JUQUEEN

- Modern Fortran, Open Source
- Distributed with MPI in tiles
- CPU-parallelized with OpenMP



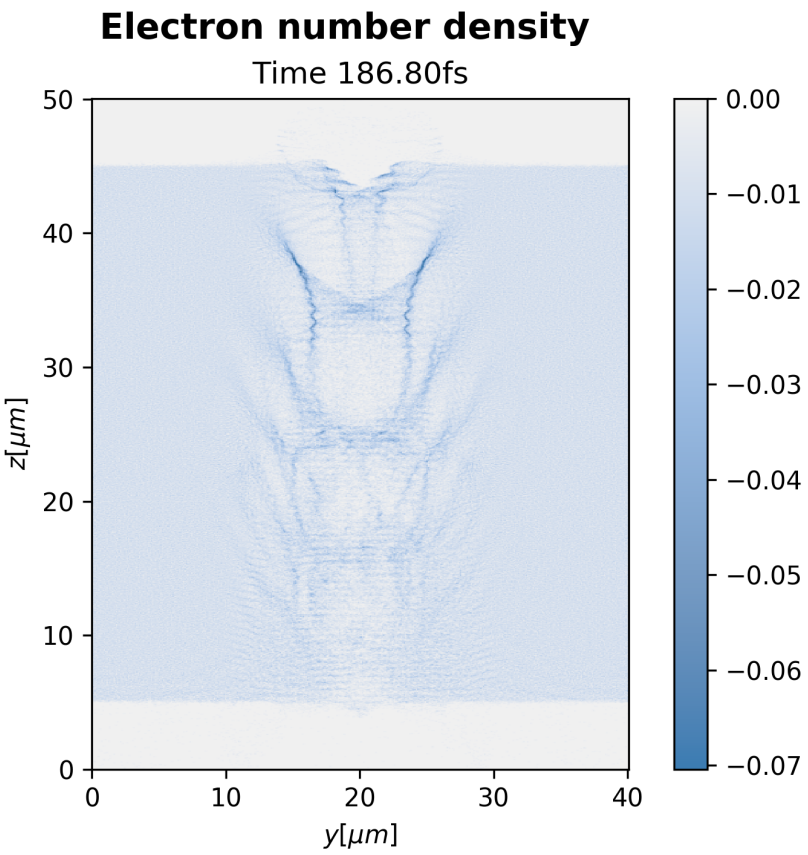
Sample Simulation

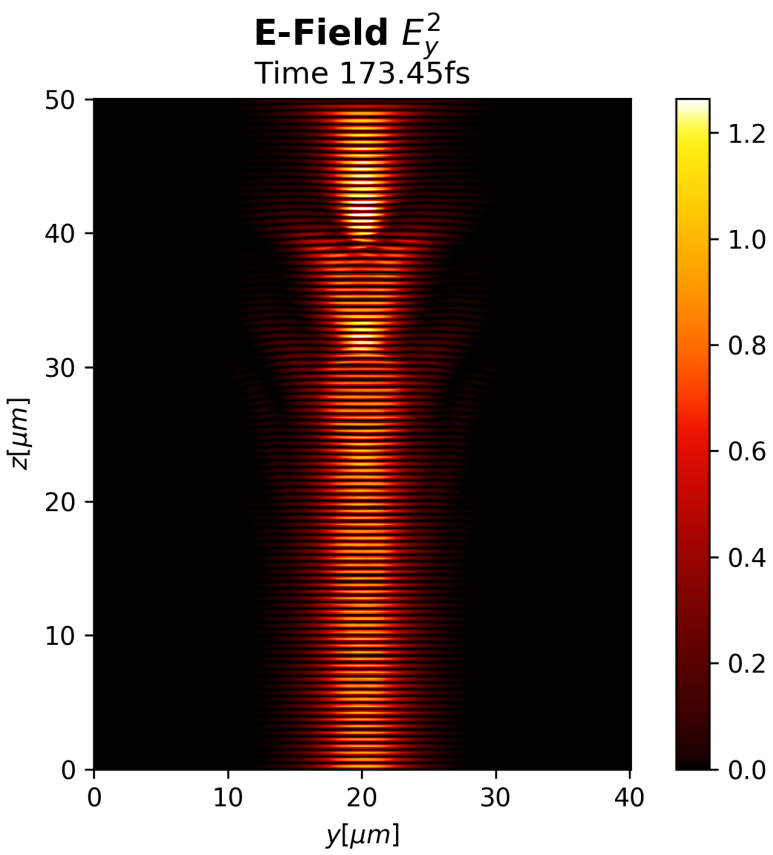
Visualizing different quantities



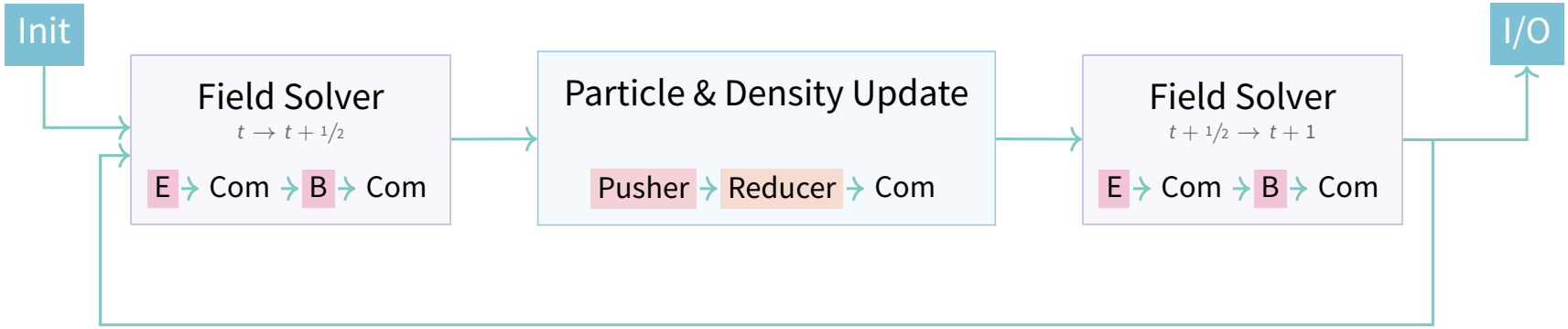
Sample Simulation

Visualizing different quantities

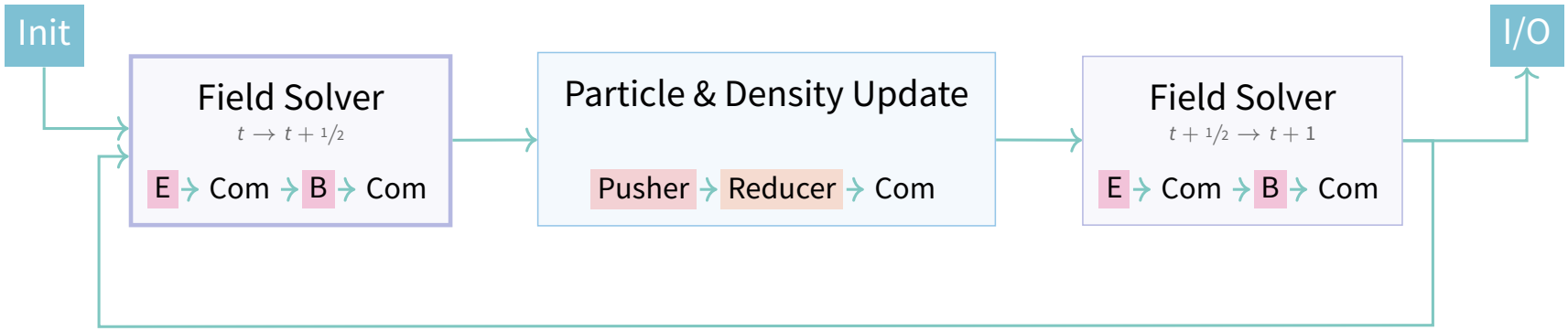




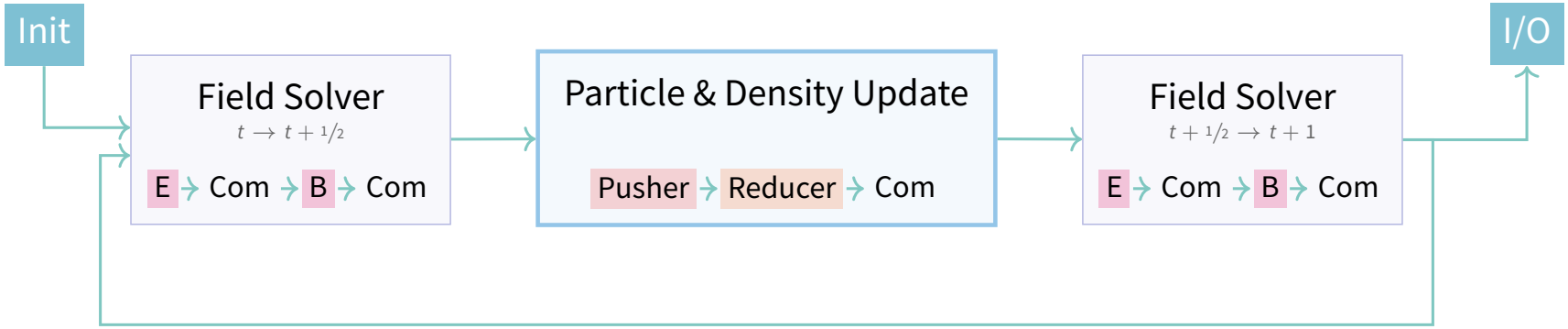
Steps of Algorithm



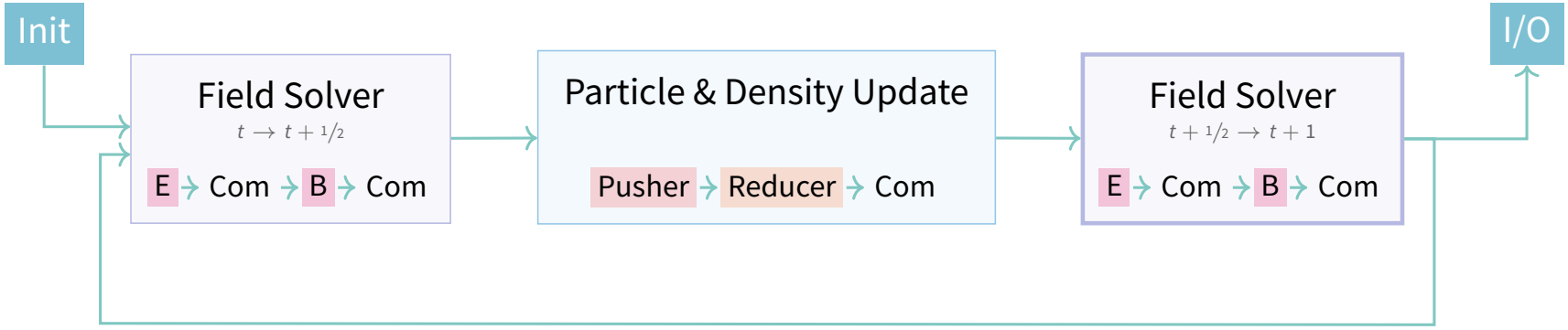
Steps of Algorithm



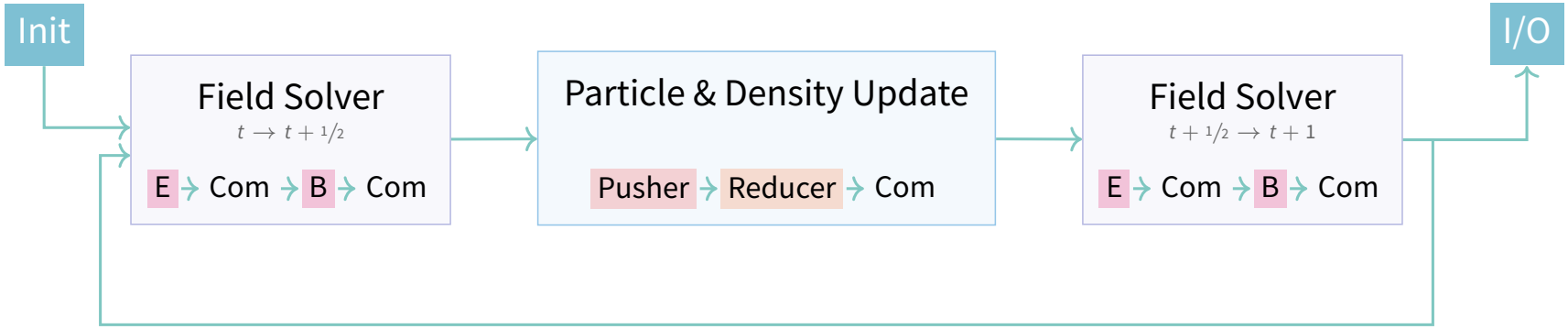
Steps of Algorithm



Steps of Algorithm



Steps of Algorithm



- E, B Already on GPU with OpenACC (small kernels)
- Pusher Focus of this paper
- Reducer Future step

Acceleration for GPUs

Acceleration for GPUs

OpenACC

OpenACC in JuSPIC

A long story

- Field solvers use OpenACC (simple code)

```
!$acc kernels loop collapse(3) present(e,b,ji)
do i3=i3mn-1,i3mx+1
   do i2=i2mn-1,i2mx+1
      do i1=i1mn-1,i1mx+1
         e(i1,i2,i3)%X=e(i1,i2,i3)%X
         ! etc
```

- Data movement with OpenACC (incl. resident parts)
- But Pusher no easy feat

OpenACC Pusher

Complicated structures

- At start of porting: `Pusher` kernel too complicated for parsing by compiler
  - Large routine (many registers)
  - Operations on whole fields (it's Fortran after all)
  - Structured data types (with alloctables)
- Long investigation to get runnable code
- Good performance complicated
- Reported in other publication (beyond scope here, appendix)

Accelerating Plasma Physics with GPUs

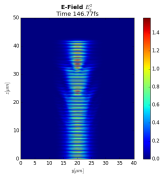
JuSPIC with OpenACC and CUDA Fortran

Andreas Herten, Dirk Pleiter, Dirk Brömmel
Jülich Supercomputing Centre

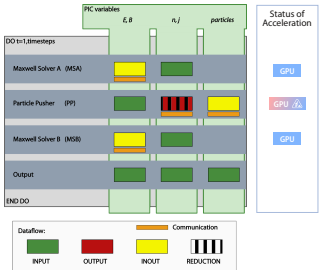
JuSPIC

Jülich Scalable Particle-in-Cell Code

- Based on plasma simulation code *PSC* (by H. Ruhl)
- 3D electromagnetic Particle-in-Cell code
- Solves relativistic Vlasov equation, coupled to Maxwell equations in finite-difference time-domain scheme
- Cartesian geometry; arbitrary number of particle species

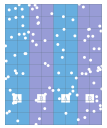


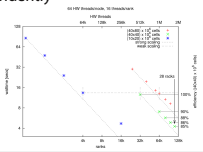
Workflow



Techniques

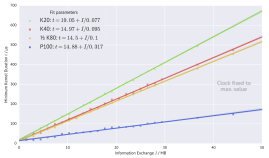
- Modern Fortran
- Fully distributed with MPI
  - Domain decomposition: tiles
- CPU-parallelized with OpenMP
  - Local decomposition: slices
  - A, B processed independently
- Scales to full JUQUEEN supercomputer





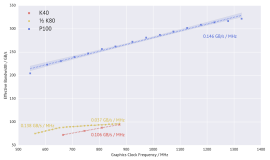
Performance Model

Based on information exchange: $t(N_{Part}) = a + l(N_{Part})/\beta$



Kernel duration vs. information exchange

- Information exchanged for kernel
  → Is lower limit of exploited bandwidth
- Effective bandwidth: K80 – 100 GB/s; P100 – 317 GB/s
- GPU kernel possibly latency-limited (many registers)



Effective bandwidths vs. clock frequencies for K40, K80, P100

- K80: Two regions (*left: performance depending on clock; right: nearly constant*)
- P100: JuSPIC benefits from new GPU design

Status of Porting and Acceleration

Particle Pusher

Three parts:
- Solve Maxwell equations with OpenACC *(not shown here)* GPU ✓
- Update of particle position & momentum (*pusher*) GPU
- Update of densities (*reduction*) GPU

CPU Version
- Single core (OpenMP disabled)
- Original data structure (linked list) temporarily moved to array
  → **overhead**

Fast OpenACC
- **Rewrite of entire computing kernel necessary!**
- Few Fortran features used (*arrays…*)

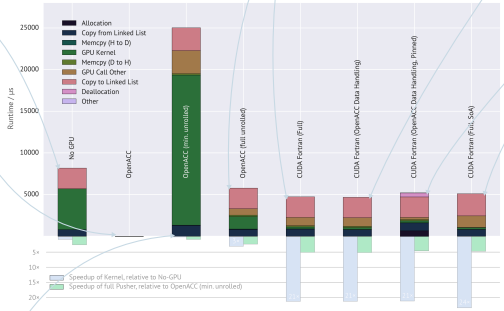
CUDA Fortran
```
call gpupusher<<<dim3(nBlocks, 1, 1), dim3(nThreads, 1, 1)>>>(...)
```
- Translation to CUDA Fortran kernel
- Helper data (scalars, 3D vectors) handled by OpenACC
- Particle pos., mom. via CUDA
- GPU-compatible through preprocessor guards

CUDA Fortran + OpenACC
```
!$acc enter data copyin(list_of_particles, nyz1, d1, dqs, ovm1, ovm2, lbounds, ubounds)
!$acc enter data use_device(list_of_particles, e, b, nyz1, d1, dqs, ovm1, ovm2, lbounds, ubounds)
```
- All data (incl. large arrays) handled by OpenACC
- Few code changes necessarry

Initial OpenACC Port
```
!$acc parallel loop private(pp,root,qi,mi,wi)
present(e,b) copy(list_of_particles)
```
- **Not running!**
- Breaks at first encounter

Working OpenACC Port
```
x_(1)=list_of_particles(i_particle)%vec(1)
x_(2)=list_of_particles(i_particle)%vec(2)
```
- Two changes necessary
  - Unroll some array operations
  - Limit number of gang/vector (***slow!***)

→ Fortran programming style and complex kernel challenging for OpenACC compiler



Speedup:
- Kernel (only compute) w/r/t CPU loop (single core)
- Full pusher (incl. all overhead) w/r/t initial OpenACC

GPU	CPU
Nvidia Tesla K40 *ECC on*	Intel Xeon E5-2650 (2 GHz) *Sandy Bridge*

FUDA + Pinned OpenACC
```
type(particle_type), dimension(:), allocatable :: list_of_particles
attributes(pinned) :: list_of_particles
```
- Pinned host data
- Faster data staging

CUDA Fortran, SoA
```
type :: possoa
  real, dimension(:), allocatable :: x, y, z, px, …
end type possoa
type(possoa) :: soa_list_of_particles
real, dimension(:), allocatable, device :: d_x, …
```
- Structure-of-Array data type for coalesced memory access
- Allocated once, resized dynamically
- Speedup single CPU: 24×

OpenACC and Fortran

- Support through PGI compiler
- Well-supported, example:
```
!$acc parallel loop
private(pp,root,qi,mi,wi)
present(e,b)
do i=1, n
  ! …
end do
!$acc end parallel
```
- JuSPIC: Many issues during parallelization
- Structured datatypes and array operations challenging for compiler
  → Many manual code adaptions
- Why not use CUDA Fortran?
- Portable with preprocessor guards!
```
#ifdef _CUDA
i = blockDim%x * blockIdx%x +
threadIdx%x
#else
do i=0,N
#endif
! …
```

Conclusion & Outlook

Conclusion
- First progress made in GPU-acceleration of JuSPIC
- Hybrid code: OpenACC and CUDA Fortran
- Changes in data layout necessary (expensive!)
- Benefit from P100 architecture

Outlook
- Reduction on GPU
- Minimization of host/devices copies
- Lowering of overhead of data layout transformations
- Evaluate data layout change for rest of JuSPIC
- Parallelization on slice / tile level
- Parallelization on multiple GPUs

JÜLICH APPLICATION LAB NVIDIA

JÜLICH FORSCHUNGSZENTRUM

OpenACC Pusher

Complicated structures

JÜLICH FORSCHUNGSZENTRUM

- At start of porting: `Pusher` kernel too complicated for parsing by compiler
  - Large routine (many registers)
  - Operations on whole fields (it's Fortran after all)
  - Structured data types (with alloctables)
- Long investigation to get runnable code
- Good performance complicated
- Reported in other publication (beyond scope here, appendix)

→ Use CUDA Fortran

Acceleration for GPUs

CUDA Fortran

Introduction to CUDA Fortran

It's like CUDA C/C++,… but for Fortran

- Available in PGI Fortran compiler
- Adds CUDA extensions to Fortran

Introduction to CUDA Fortran

It's like CUDA C/C++,… but for Fortran

- Available in PGI Fortran compiler
- Adds CUDA extensions to Fortran
- Examples (from JuSPIC):

Introduction to CUDA Fortran

It's like CUDA C/C++,… but for Fortran

- Available in PGI Fortran compiler
- Adds CUDA extensions to Fortran
- Examples (from JuSPIC):
  - Define device function along-side host function

```
type(particle_type), dimension(slice(1)%n) ::
↪  list_of_particles, list_of_particles_d
attributes(device) :: list_of_particles_d
```

Introduction to CUDA Fortran

It's like CUDA C/C++,… but for Fortran

- Available in PGI Fortran compiler
- Adds CUDA extensions to Fortran
- Examples (from JuSPIC):
  - Define device function along-side host function

```
type(particle_type), dimension(slice(1)%n) ::
↪  list_of_particles, list_of_particles_d
attributes(device) :: list_of_particles_d
```

  - Copy to device

```
list_of_particles_d = list_of_particles
```

Introduction to CUDA Fortran

It's like CUDA C/C++,… but for Fortran

- Available in PGI Fortran compiler
- Adds CUDA extensions to Fortran
- Examples (from JuSPIC):
  - Define device function along-side host function

```
type(particle_type), dimension(slice(1)%n) ::
↪  list_of_particles, list_of_particles_d
attributes(device) :: list_of_particles_d
```

  - Copy to device

```
list_of_particles_d = list_of_particles
```

  - Define kernel

```
attributes(global) subroutine gpupusher(list_of_particles, ...)
```

Introduction to CUDA Fortran

It's like CUDA C/C++,… but for Fortran

- Available in PGI Fortran compiler
- Adds CUDA extensions to Fortran
- Examples (from JuSPIC):
  - Define device function along-side host function

```
type(particle_type), dimension(slice(1)%n) ::
↪  list_of_particles, list_of_particles_d
attributes(device) :: list_of_particles_d
```

  - Copy to device

```
list_of_particles_d = list_of_particles
```

  - Define kernel

```
attributes(global) subroutine gpupusher(list_of_particles, ...)
```

  - Call kernel

```
call gpupusher<<<dim3(nBlocks, 1, 1), dim3(nThreads, 1,
↪  1)>>>(list_of_particles_d, ...)
```

CUDA Fortran Portability

Not as portable as OpenACC, but it's alright

- CUDA Fortran: more powerful approach
- Portability suffers…

CUDA Fortran Portability

Not as portable as OpenACC, but it's alright

- CUDA Fortran: more powerful approach
- Portability suffers…
- … but can be mitigated!

1. Use OpenACC as much as possible, e.g. for data movements
   OpenACC mixes well together with CUDA Fortran

```
!$acc enter data copyin(list_of_particles, ...)
```

CUDA Fortran Portability

Not as portable as OpenACC, but it's alright

- CUDA Fortran: more powerful approach
- Portability suffers…
- … but can be mitigated!

1. Use OpenACC as much as possible, e.g. for data movements
   OpenACC mixes well together with CUDA Fortran

```
!$acc enter data copyin(list_of_particles, ...)
```

2. Use pre-processor directives for rest

```
#ifdef _CUDA
  i = blockDim%x * (blockIdx%x - 1) + threadIdx%x
#else
  do i = lbound(a, 1), ubound(a, 1)
#endif
```

Acceleration for GPUs

Data Layout Analysis

Strategies for Data Layout

Because data is not solely data

JÜLICH FORSCHUNGSZENTRUM

- Benchmark different data layouts and transfer strategies
- Sub-parts of Pusher:

$\sum$	*Everything*
Allocate	Allocate host-side data structures
LL2F	Convert linked-list data structure to field
H2D	Copy data from host to device
Kernel	Run kernel
D2H	Copy data from device to host
Other	Left-over time (synchronization, etc.)
F2LL	Copy flat field back to linked list

- Benchmarking on **JURON**

Data Layout Experiments

Description of experiments

Initial All particles stored in single field, one particle after another; data copied to/from GPU with Fortran *(baseline)*

Data Layout Experiments

Description of experiments

JÜLICH FORSCHUNGSZENTRUM

Initial All particles stored in single field, one particle after another; data copied to/from GPU with Fortran *(baseline)*

in µs	$\sum$	Allocate	LL2F	H2D	Kernel	D2H	Others	F2LL
Initial	8040	–	567	82	84	62	350	6885

Data Layout Experiments

Description of experiments

JÜLICH FORSCHUNGSZENTRUM

Initial All particles stored in single field, one particle after another; data copied to/from GPU with Fortran *(baseline)*

Exp 1 As *Initial*, but data copied with OpenACC *copy* directives

in µs	$\sum$	Allocate	LL2F	H2D	Kernel	D2H	Others	F2LL
Initial	8040	–	567	82	84	62	350	6885
Exp 1	10435	–	353	80	82	91	380	9440

Data Layout Experiments

Description of experiments

JÜLICH FORSCHUNGSZENTRUM

Initial All particles stored in single field, one particle after another; data copied to/from GPU with Fortran *(baseline)*

Exp 1 As *Initial*, but data copied with OpenACC *copy* directives

Exp 2 As *Exp 1*, but data copied from pinned host memory

in µs	$\sum$	Allocate	LL2F	H2D	Kernel	D2H	Others	F2LL
Initial	8040	–	567	82	84	62	350	6885
Exp 1	10435	–	353	80	82	91	380	9440
Exp 2	9695	564	527	79	83	72	108	7973

Data Layout Experiments

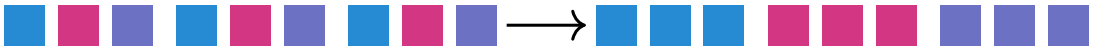
Description of experiments

Initial All particles stored in single field, one particle after another; data copied to/from GPU with Fortran *(baseline)*

Exp 1 As *Initial*, but data copied with OpenACC *copy* directives

Exp 2 As *Exp 1*, but data copied from pinned host memory

SoA Data copied with Fortran, but instead of one field with all particle data, one field for each spatial and momentum component for particles



in µs	$\sum$	Allocate	LL2F	H2D	Kernel	D2H	Others	F2LL
Initial	8040	–	567	82	84	62	350	6885
Exp 1	10435	–	353	80	82	91	380	9440
Exp 2	9695	564	527	79	83	72	108	7973
SoA	7811	1	844	66	77	53	376	6386

Data Layout Experiments

Discussion of results

in µs	$\sum$	Allocate	LL2F	H2D	Kernel	D2H	Others	F2LL
Initial	8040	–	567	82	84	62	350	6885
Exp 1	10435	–	353	80	82	91	380	9440
Exp 2	9695	564	527	79	83	72	108	7973
SoA	7811	1	844	66	77	53	376	6386

Data Layout Experiments

Discussion of results

in µs	$\sum$	Allocate	LL2F	H2D	Kernel	D2H	Others	F2LL
Initial	8040	–	567	82	84	62	350	6885
Exp 1	10435	–	353	80	82	91	380	9440
Exp 2	9695	564	527	79	83	72	108	7973
SoA	7811	1	844	**66**	**77**	**53**	376	6386

- **SoA**: fastest, looking (also) at raw GPU runtimes

Data Layout Experiments

Discussion of results

in µs	$\sum$	Allocate	LL2F	H2D	Kernel	D2H	Others	F2LL
Initial	8040	–	567	82	84	62	350	6885
Exp 1	10435	–	353	80	82	91	380	9440
Exp 2	9695	564	527	79	83	72	108	7973
SoA	7811	1	**844**	66	77	53	376	6386

- **SoA**: fastest, looking (also) at raw GPU runtimes – but slowest for change of data structures (six fields vs. one)

Data Layout Experiments

Discussion of results

in µs	$\sum$	Allocate	LL2F	H2D	Kernel	D2H	Others	F2LL
Initial	8040	–	567	82	84	62	350	6885
Exp 1	10435	–	353	80	82	91	380	9440
Exp 2	9695	564	527	79	83	72	**108**	7973
SoA	7811	1	844	66	77	53	376	6386

- SoA: fastest, looking (also) at raw GPU runtimes – but slowest for change of data structures (six fields vs. one)
- **Exp 2**: least overhead; pinned memory allows for direct data access

Data Layout Experiments

Discussion of results

in µs	$\sum$	Allocate	LL2F	H2D	Kernel	D2H	Others	F2LL
Initial	8040	–	567	82	84	62	350	6885
Exp 1	10435	–	353	80	82	91	380	9440
Exp 2	9695	**564**	527	79	83	**72**	**108**	7973
SoA	7811	1	844	66	77	53	376	6386

- SoA: fastest, looking (also) at raw GPU runtimes – but slowest for change of data structures (six fields vs. one)
- **Exp 2**: least overhead; pinned memory allows for direct data access – but allocation overhead is not fully resolved

Data Layout Experiments

Discussion of results

in µs	$\sum$	Allocate	LL2F	H2D	Kernel	D2H	Others	F2LL
Initial	8040	–	567	82	84	62	350	6885
Exp 1	10435	–	353	80	82	91	380	**9440**
Exp 2	9695	564	527	79	83	72	108	7973
SoA	7811	1	844	66	77	53	376	6386

- SoA: fastest, looking (also) at raw GPU runtimes – but slowest for change of data structures (six fields vs. one)
- Exp 2: least overhead; pinned memory allows for direct data access – but allocation overhead is not fully resolved
- **Exp 1**: also ok for raw GPU times, but large F2LL overhead *(more on that later)*

Data Layout Experiments

Architecture Comparison

in µs	$\sum$	Allocate	LL2F	H2D	Kernel	D2H	Others	F2LL
				JURON				
Initial	8040	–	567	82	84	62	350	6885
Exp 1	10435	–	353	80	82	91	380	9440
Exp 2	9695	564	527	79	83	72	108	7973
SoA	7811	1	844	66	77	53	376	6386
				JUHYDRA				
Initial	4956	–	908	267	229	208	736	2600
Exp 1	4687	–	764	232	229	198	804	2455
Exp 2	5328	577	1027	224	230	192	23	2651
SoA	4880	1	786	204	208	173	827	2674

Data Layout Experiments

Architecture Comparison

in µs	$\sum$	Allocate	LL2F	H2D	Kernel	D2H	Others	F2LL
JURON								
Initial	8040	–	567	82	84	62	350	6885
Exp 1	10435	–	353	80	82	91	380	9440
Exp 2	9695	564	527	79	83	72	108	7973
SoA	7811	1	844	66	77	53	376	6386
JUHYDRA								
Initial	4956	–	908	267	229	208	736	2600
Exp 1	4687	–	764	232	229	198	804	2455
Exp 2	5328	577	1027	224	230	192	23	2651
SoA	4880	1	786	204	208	173	827	2674

2.8×

Data Layout Experiments

Architecture Comparison

in µs	$\sum$	Allocate	LL2F	H2D	Kernel	D2H	Others	F2LL
JURON								
Initial	8040	–	567	82	84	62	350	6885
Exp 1	10435	–	353	80	82	91	380	9440
Exp 2	9695	564	527	79	83	72	108	7973
SoA	7811	1	844	66	77	53	376	6386
JUHYDRA								
Initial	4956	–	908	267	229	208	736	2600
Exp 1	4687	–	764	232	229	198	804	2455
Exp 2	5328	577	1027	224	230	192	23	2651
SoA	4880	1	786	204	208	173	827	2674

3.2×

2.8×

Data Layout Experiments

Architecture Comparison

in µs	$\sum$	Allocate	LL2F	H2D	Kernel	D2H	Others	F2LL
JURON								
Initial	8040	–	567	82	84	62	350	6885
Exp 1	10435	–	353	80	82	91	380	9440
Exp 2	9695	564	527	79	83	72	108	7973
SoA	7811	1	844	66	77	53	376	6386
JUHYDRA								
Initial	4956	–	908	267	229	208	736	2600
Exp 1	4687	–	764	232	229	198	804	2455
Exp 2	5328	577	1027	224	230	192	23	2651
SoA	4880	1	786	204	208	173	827	2674

0.6×

3.2×

2.8×

Data Layout Experiments

Architecture Comparison

JÜLICH FORSCHUNGSZENTRUM

in µs	$\sum$	Allocate	LL2F	H2D	Kernel	D2H	Others	F2LL
JURON								
Initial	8040	–	567	82	84	62	350	6885
Exp 1	10435	–	353	80	82	91	380	9440
Exp 2	9695	564	527	79	83	72	108	7973
SoA	7811	1	844	66	77	53	376	6386
JUHYDRA								
Initial	4956	–	908	267	229	208	736	2600
Exp 1	4687	–	764	232	229	198	804	2455
Exp 2	5328	577	1027	224	230	192	23	2651
SoA	4880	1	786	204	208	173	827	2674

3.2×

0.6×

2.8×

0.3×

Data Layout Experiments

Architecture Comparison

JÜLICH FORSCHUNGSZENTRUM

in µs	$\sum$	Allocate	LL2F	H2D	Kernel	D2H	Others	F2LL
JURON								
Initial	8040	–	567	82	84	62	350	6885
Exp 1	10435				82	91	380	9440
Exp 2	9695				83	72	108	7973
SoA	7811				77	53	376	6386
JUHYDRA								
Initial	4956	–	908	267	229	208	736	2600
Exp 1	4687	–	764	232	229	198	804	2455
Exp 2	5328	577	1027	224	230	192	23	2651
SoA	4880	1	786	204	208	173	827	2674

Why!?

2×

0.6×

2.8×

0.3×

Acceleration for GPUs

Data Layout Conversion

Conversion of Data Layouts

Why is F2LL so slow?

- Parts of F2LL
  - Kill old linked list of particles[1]
  - Initialize new, empty linked list of particles
  - Loop through field(s) of particle information…
  - … add each particle to linked list, update pointers

[1] Start with first particle, progress along links, remove each particle

Conversion of Data Layouts

Why is F2LL so slow?

- Parts of F2LL
  - Kill old linked list of particles[1]
  - Initialize new, empty linked list of particles
  - Loop through field(s) of particle information…
  - … add each particle to linked list, update pointers
- `add_one_to_list`

```
allocate(list%tail%next)
nullify(list%tail%next%next)
list%tail%next%particle = particle
list%tail => list%tail%next
```

[1] Start with first particle, progress along links, remove each particle

Conversion of Data Layouts

Why is F2LL so slow?

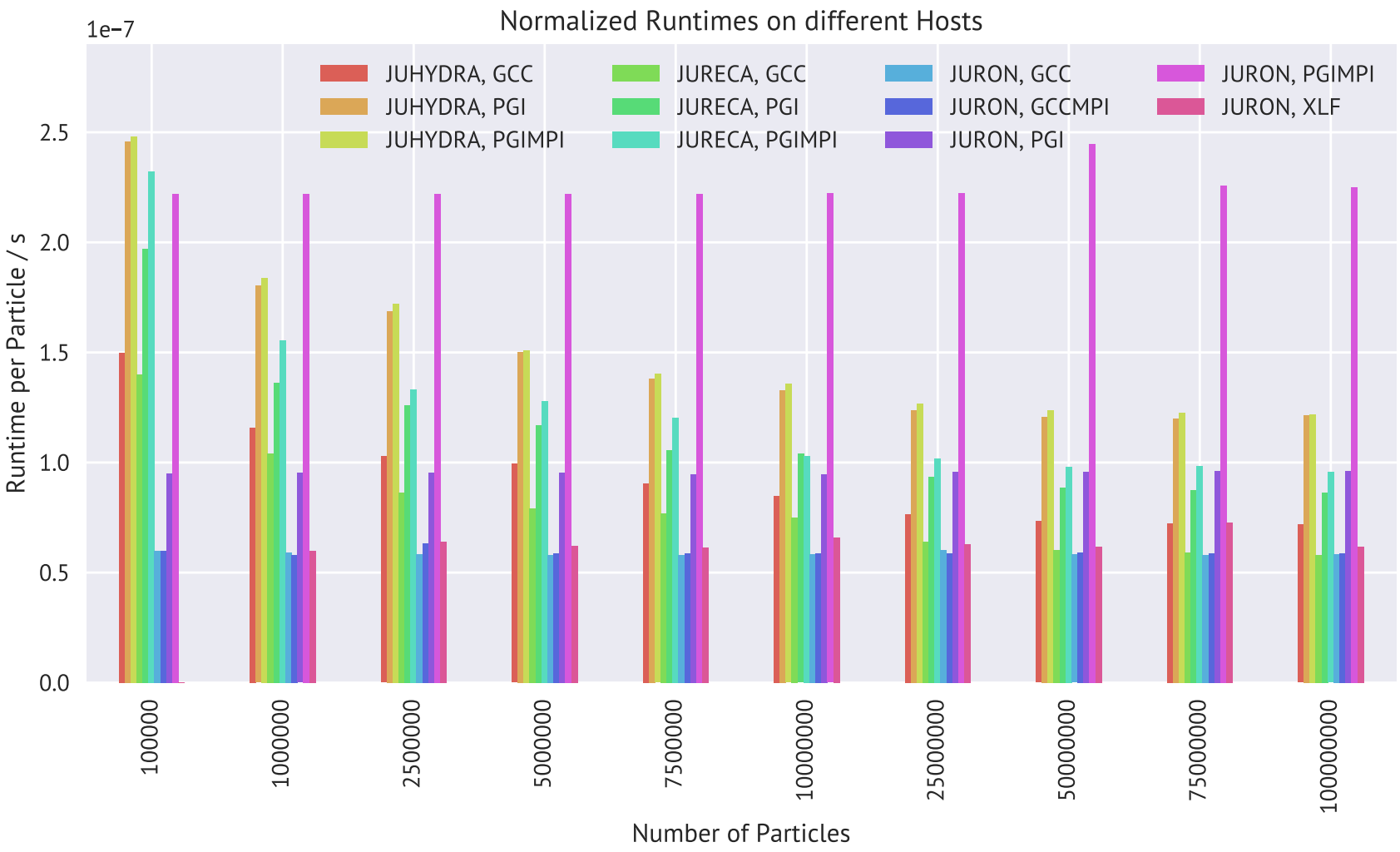
- Parts of F2LL
  - Kill old linked list of particles[1]
  - Initialize new, empty linked list of particles
  - Loop through field(s) of particle information…
  - … add each particle to linked list, update pointers
- `add_one_to_list`

```
allocate(list%tail%next)
nullify(list%tail%next%next)
list%tail%next%particle = particle
list%tail => list%tail%next
```

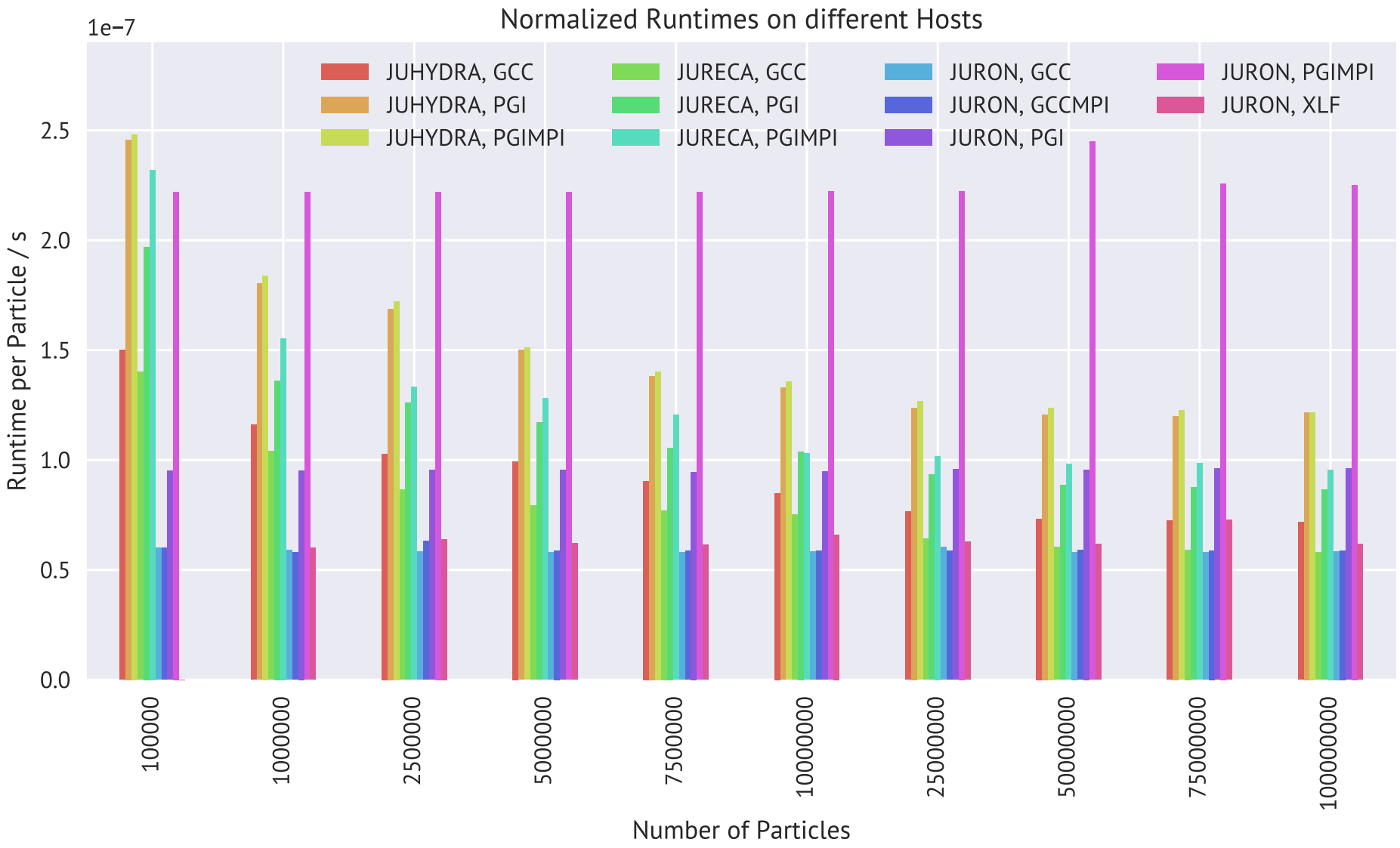
⇒ ***Benchmark***

[1] Start with first particle, progress along links, remove each particle

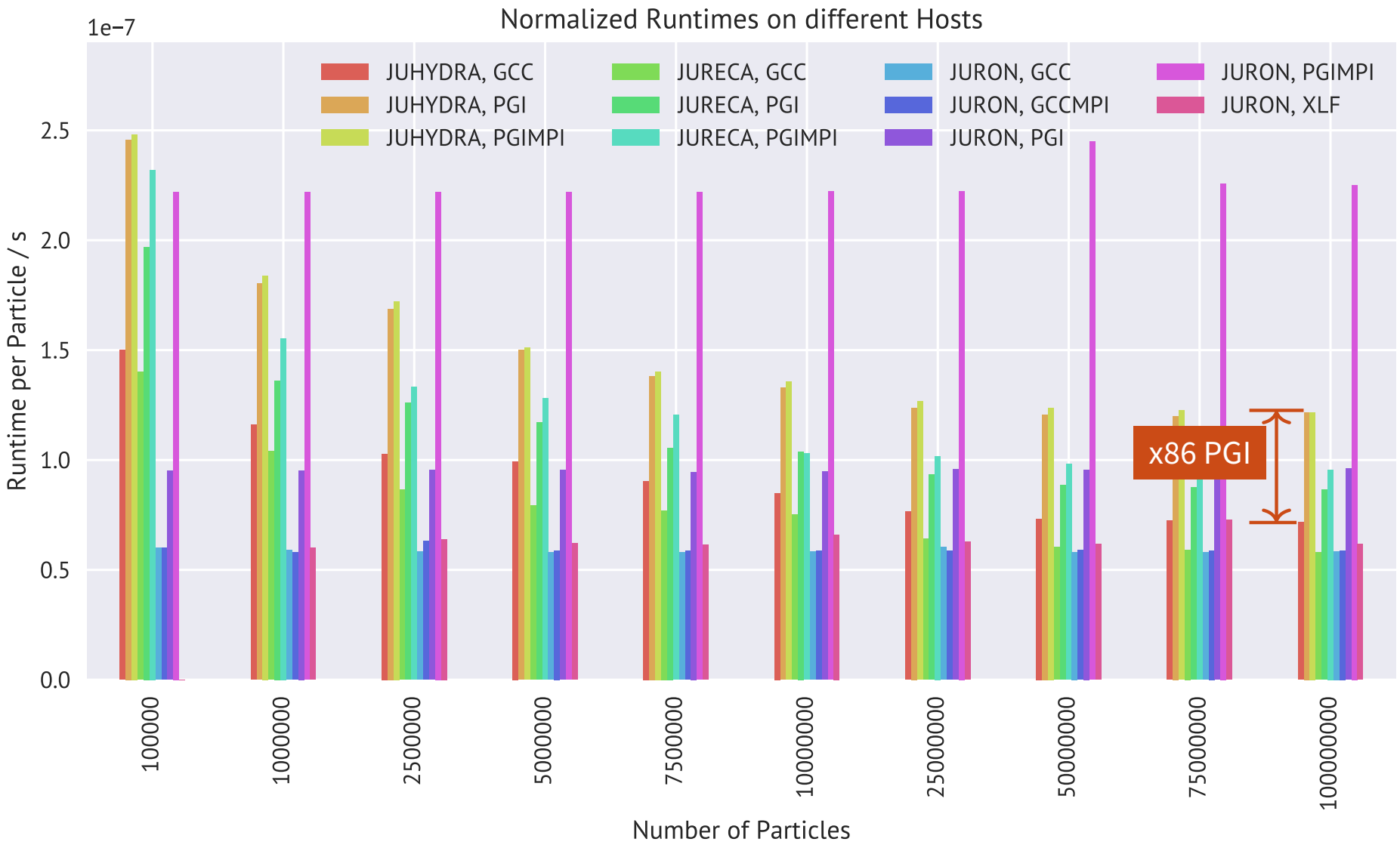
Compiler/MPI Timings



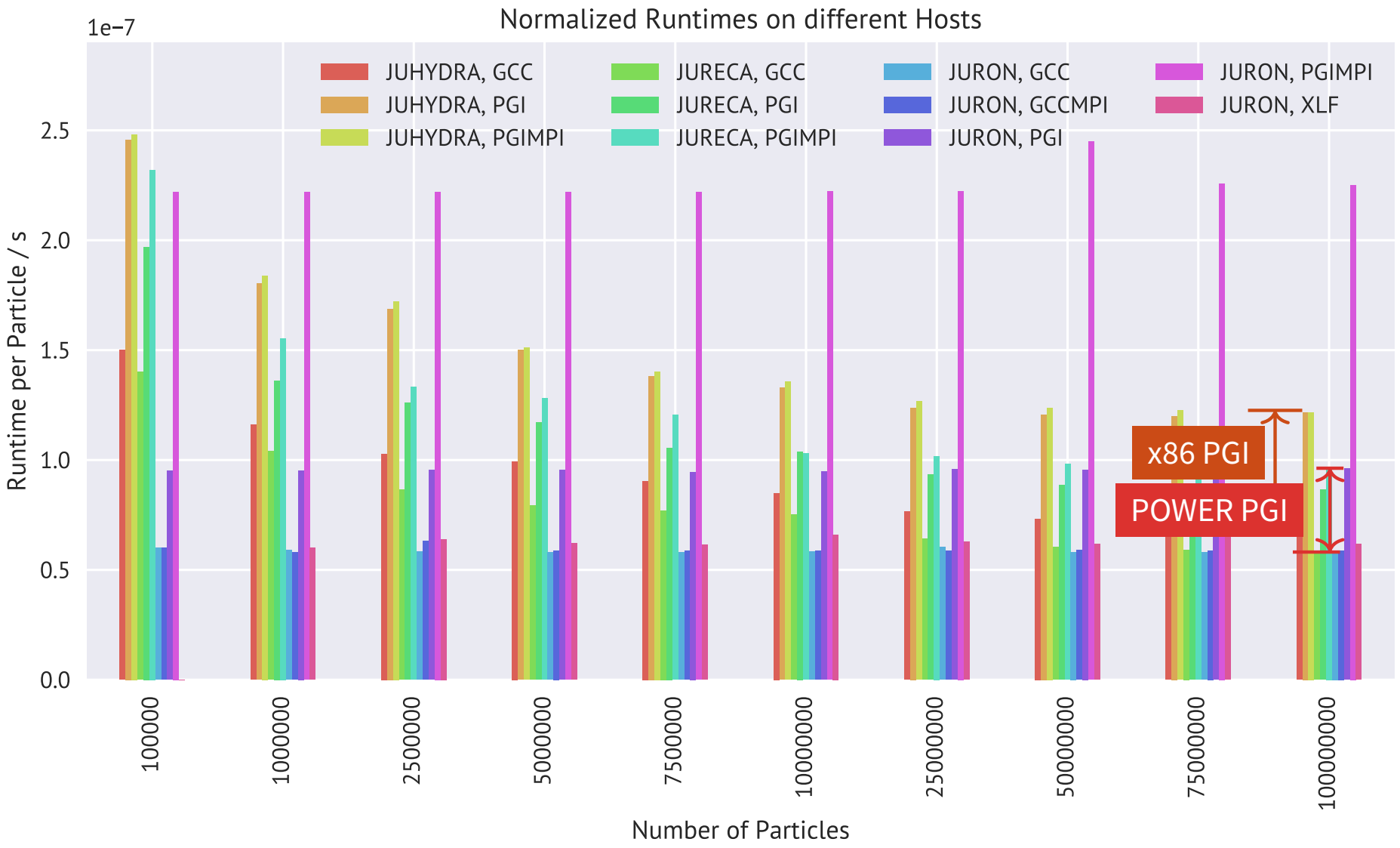
Compiler/MPI Timings



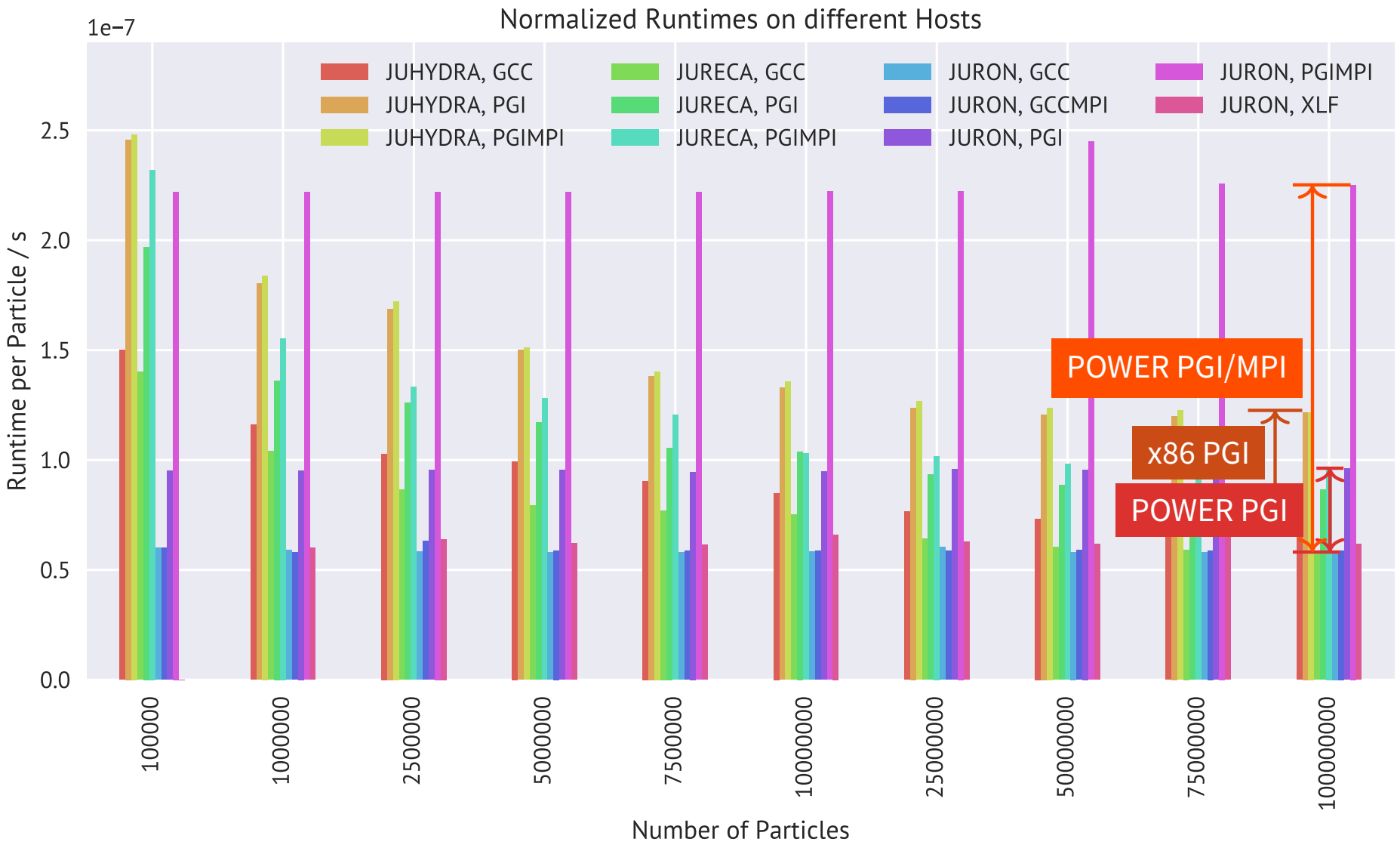
Compiler/MPI Timings



Compiler/MPI Timings



Compiler/MPI Timings



Compiler Investigation

Is MPI Slow? And, by the way, which MPI!?

- *PGIMPI*: MPI version shipped with PGI
- Not actively used in GPU version of JuSPIC, but in future

Compiler Investigation

Is MPI Slow? And, by the way, which MPI!?

- *PGIMPI*: MPI version shipped with PGI
- Not actively used in GPU version of JuSPIC, but in future
- `add_one_to_list` benchmark does not use MPI at all!
  Replacing `pgfortran` by `mpifort` leads to performance decrease

Compiler Investigation

Is MPI Slow? And, by the way, which MPI!?

- *PGIMPI*: MPI version shipped with PGI
- Not actively used in GPU version of JuSPIC, but in future
- `add_one_to_list` benchmark does not use MPI at all!
  Replacing `pgfortran` by `mpifort` leads to performance decrease

→ ***Benchmark*** compilers – with PAPI [3] instrumentation

Compiler Investigation

Is MPI Slow? And, by the way, which MPI!?

- *PGIMPI*: MPI version shipped with PGI
- Not actively used in GPU version of JuSPIC, but in future
- `add_one_to_list` benchmark does not use MPI at all!
  Replacing `pgfortran` by `mpifort` leads to performance decrease

→ ***Benchmark*** compilers – with PAPI [3] instrumentation

System	JURON						JUHYDRA	
Compiler	GCC	GCCMPI	PGI	PGIMPI	PGIMPI*	XLF	PGI	PGIMPI
Time pP/ns	36	37	46	154	48	41	32	32
Instructions pP	121	121	243	462	243	121	210	210

See appendix for some more counters

Further Investigation/Mitigation

- MPI version shipped with PGI on POWER is slow, because it issues many instructions

Further Investigation/Mitigation

- MPI version shipped with PGI on POWER is slow, because it issues many instructions
- Further study: Identical assembly code generated as MPI-less version…

Further Investigation/Mitigation

- MPI version shipped with PGI on POWER is slow, because it issues many instructions
- Further study: Identical assembly code generated as MPI-less version…
- … but includes call to `malloc()`!

Further Investigation/Mitigation

JÜLICH FORSCHUNGSZENTRUM

- MPI version shipped with PGI on POWER is slow, because it issues many instructions
- Further study: Identical assembly code generated as MPI-less version…
- … but includes call to `malloc()`!
- Different libraries linked for PGI and PGIMPI cases!

Further Investigation/Mitigation

- MPI version shipped with PGI on POWER is slow, because it issues many instructions
- Further study: Identical assembly code generated as MPI-less version…
- … but includes call to `malloc()`!
- Different libraries linked for PGI and PGIMPI cases!
- `LD_PRELOAD=/lib64/libc.so.6` solves problem!

Further Investigation/Mitigation

JÜLICH FORSCHUNGSZENTRUM

- MPI version shipped with PGI on POWER is slow, because it issues many instructions
- Further study: Identical assembly code generated as MPI-less version…
- … but includes call to `malloc()`!
- Different libraries linked for PGI and PGIMPI cases!
- `LD_PRELOAD=/lib64/libc.so.6` solves problem!

⇒ Slow MPI-aware `malloc()`?

Further Investigation/Mitigation

- MPI version shipped with PGI on POWER is slow, because it issues many instructions
- Further study: Identical assembly code generated as MPI-less version…
- … but includes call to `malloc()`!
- Different libraries linked for PGI and PGIMPI cases!
- `LD_PRELOAD=/lib64/libc.so.6` solves problem!

⇒ Slow MPI-aware `malloc()`?

- Mitigation
  - **Bug reported**
  - For now: consider as *anomalous* overhead

Performance Modelling

Effective Bandwidth

Defining the model

- **Goal**: Compare different GPU architectures; understand behavior of JuSPIC
- Model based on **information exchanged** of GPU kernel
  - Amount of exchanged information for given number of particles
  - Time for exchange

Effective Bandwidth

Defining the model

- **Goal**: Compare different GPU architectures; understand behavior of JuSPIC
- Model based on **information exchanged** of GPU kernel
  - Amount of exchanged information for given number of particles
  - Time for exchange

$$t(N_{\text{part}}) = \alpha + I(N_{\text{part}})/\beta \,,$$

Effective Bandwidth

Defining the model

- **Goal**: Compare different GPU architectures; understand behavior of JuSPIC
- Model based on **information exchanged** of GPU kernel
  - Amount of exchanged information for given number of particles
  - Time for exchange

$$t(N_{\text{part}}) = \alpha + I(N_{\text{part}})/\beta\,,$$

N_{part} Number of particles processed

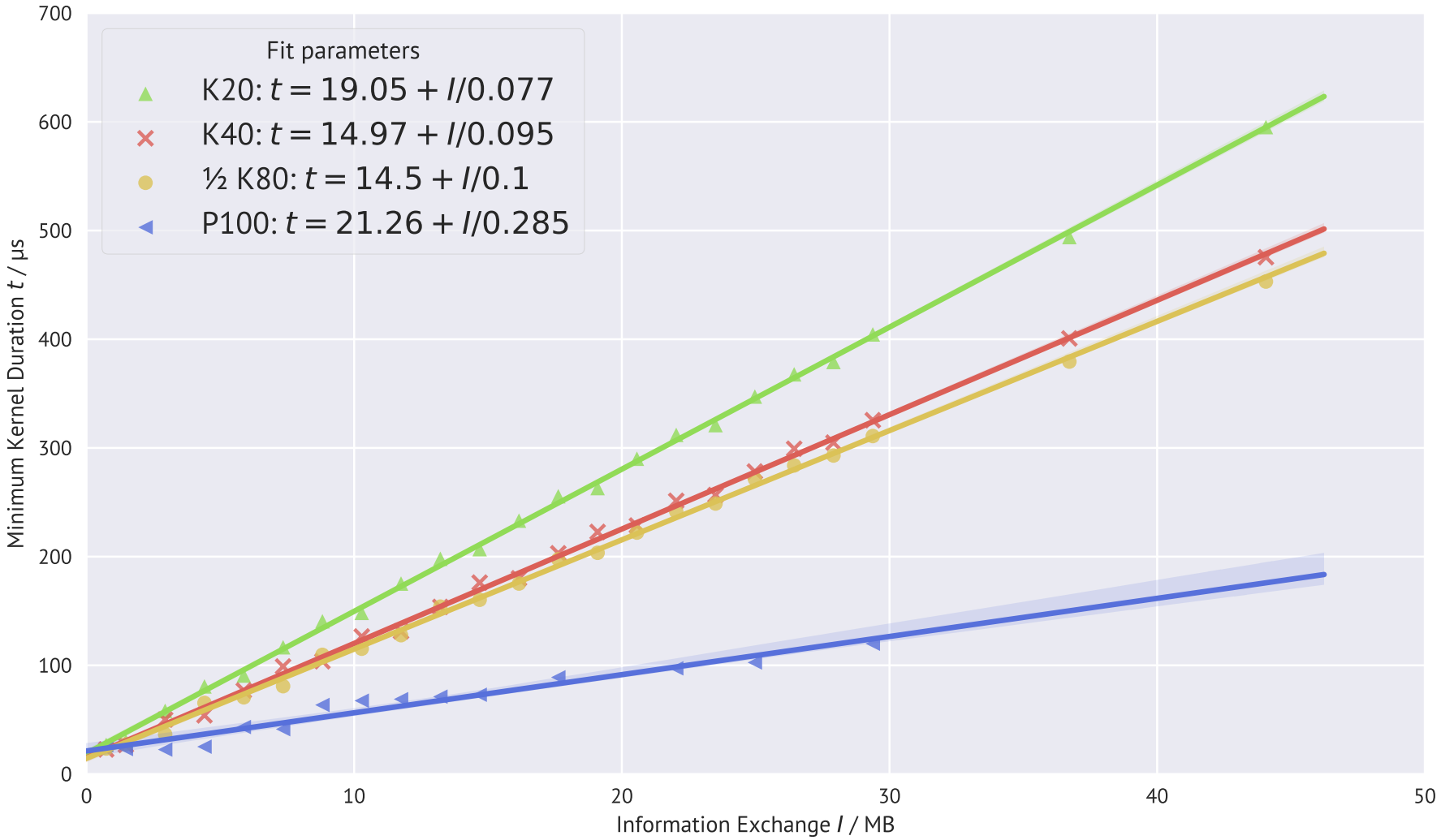
I Information exchanged (572 B (read) + 40 B (write))

t Kernel runtime

α, β Fit parameters; β: *effective bandwidth*

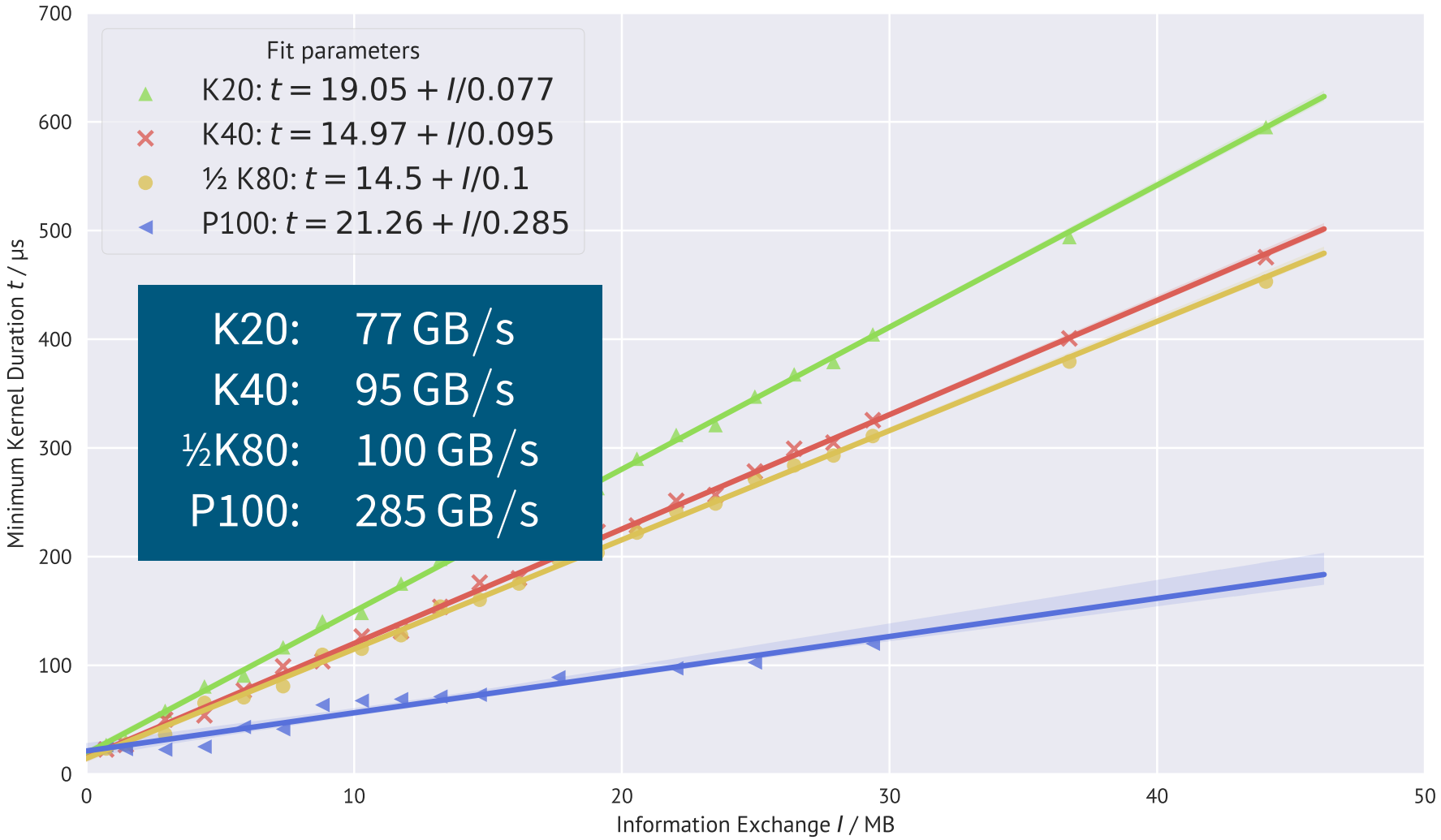
Effective Bandwidth

Measurements



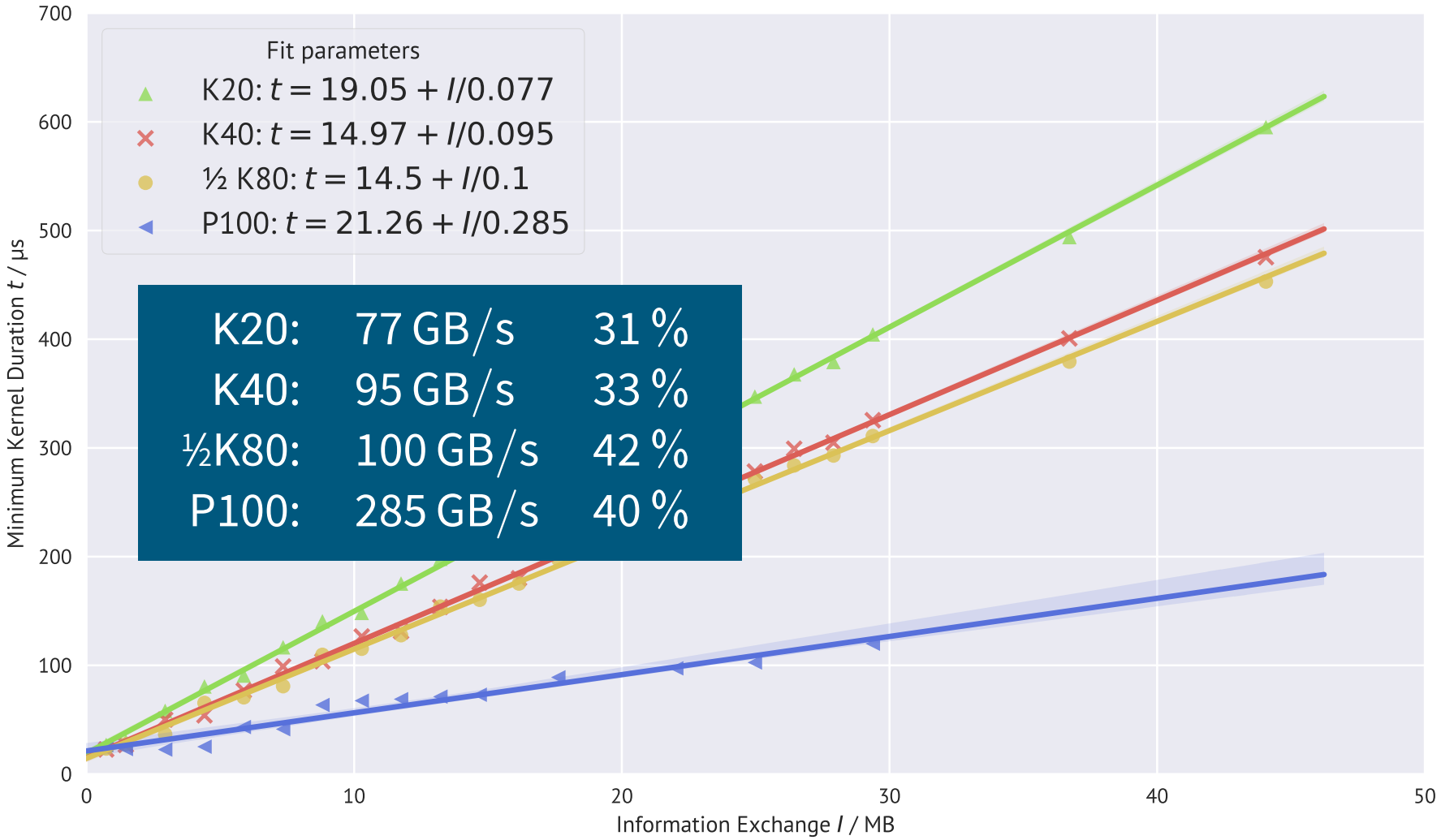
Effective Bandwidth

Measurements



Effective Bandwidth

Measurements



Clock Dependency

Defining the relation

- Another free parameter: **GPU clock rates**
- Varies significantly across GPU architecture generations and models

→ Incorporate clock into performance model

Clock Dependency

Defining the relation

- Another free parameter: **GPU clock rates**
- Varies significantly across GPU architecture generations and models

→ Incorporate clock into performance model

$$\beta(\mathcal{C}) = \gamma + \delta\,\mathcal{C}$$

Clock Dependency

Defining the relation

- Another free parameter: **GPU clock rates**
- Varies significantly across GPU architecture generations and models

→ Incorporate clock into performance model

$$\beta(\mathcal{C}) = \gamma + \delta\,\mathcal{C}$$

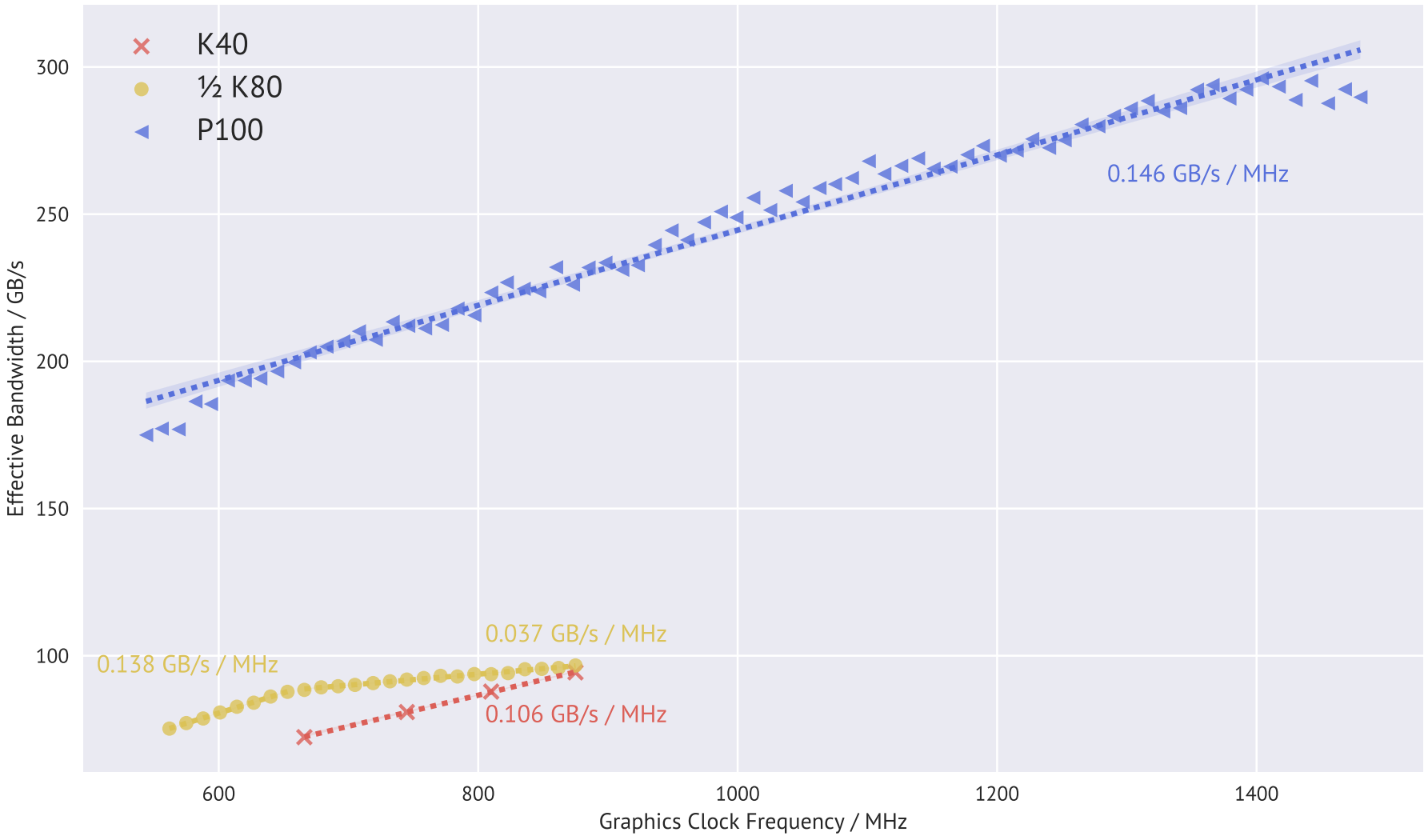
$\mathcal{C}$ GPU clock rate

β Effective bandwidth (from before)

γ, δ Fit parameters

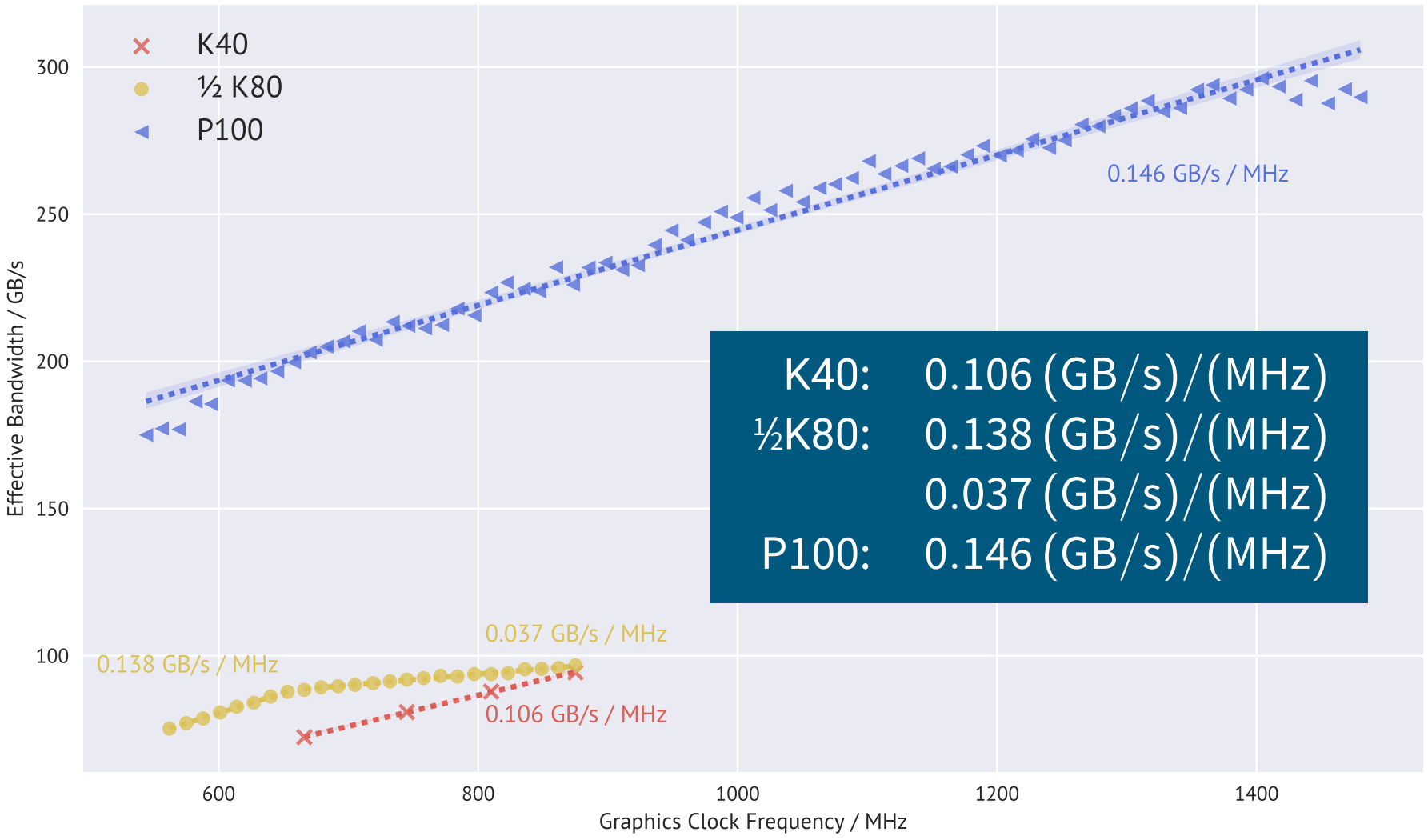
Clock Dependency

Measurements



Clock Dependency

Measurements



K40:	0.106 (GB/s)/(MHz)
½K80:	0.138 (GB/s)/(MHz)
	0.037 (GB/s)/(MHz)
P100:	0.146 (GB/s)/(MHz)

Summary, Conclusion

JÜLICH FORSCHUNGSZENTRUM

Summary

- Enabled JuSPIC for **GPU** with OpenACC & CUDA Fortran
- Particle data layout: **SoA fastest**
- **Slow memory allocation** with PGI+MPI on POWER $\rightarrow$ bug filed
- Performance model: **Information exchange** (P100: 285 GB/s)
- Studied model with **different clock rates** – P100 most efficient scaling

Future

- Port also Reducer to GPU
- Enable MPI again
- Alternatives to linked list

Summary, Conclusion

JÜLICH FORSCHUNGSZENTRUM

Summary

- Enabled **JuSPIC** for **GPU** with OpenACC & CUDA Fortran
- Particle data layout: **SoA fastest**
- **Slow memory allocation** with PGI+MPI on POWER $\rightarrow$ bug filed
- Performance model: **Information exchange** (P100: 285 GB/s)
- Studied model with **different clock rates** – P100 most efficient scaling

Future

- Port also Reducer to GPU
- Enable MPI again
- Alternatives to linked list

Thank you for your attention!

a.herten@fz-juelich.de

Appendix

Acknowledgements

- The work was done in context of two groups:

  POWER Acceleration and Design Centre A collaboration of IBM, NVIDIA, and Forschungszentrum Jülich

  NVIDIA Application Lab A collaboration of NVIDIA and Forschungszentrum Jülich
- Many thanks to Jiri Kraus from NVIDIA, who helped tremendously along the way
- JURON, a prototype system for the Human Brain Project, received co-funding from the European Union (Grant Agreement No. 604102)

Related Work

- Selection of other GPU PiC codes

  PSC The code JuSPIC is based on has been reimplemented in C and ported to GPU [4]

  PIConGPU PiC code specifically developed for GPUs [5]
- Minsky porting experiences
  - "Addressing Materials Science Challenges Using GPU-accelerated POWER8 Nodes" [6]
  - "A Performance Model for GPU-Accelerated FDTD Applications" [7]
- … more in paper!

OpenACC Performance Progression

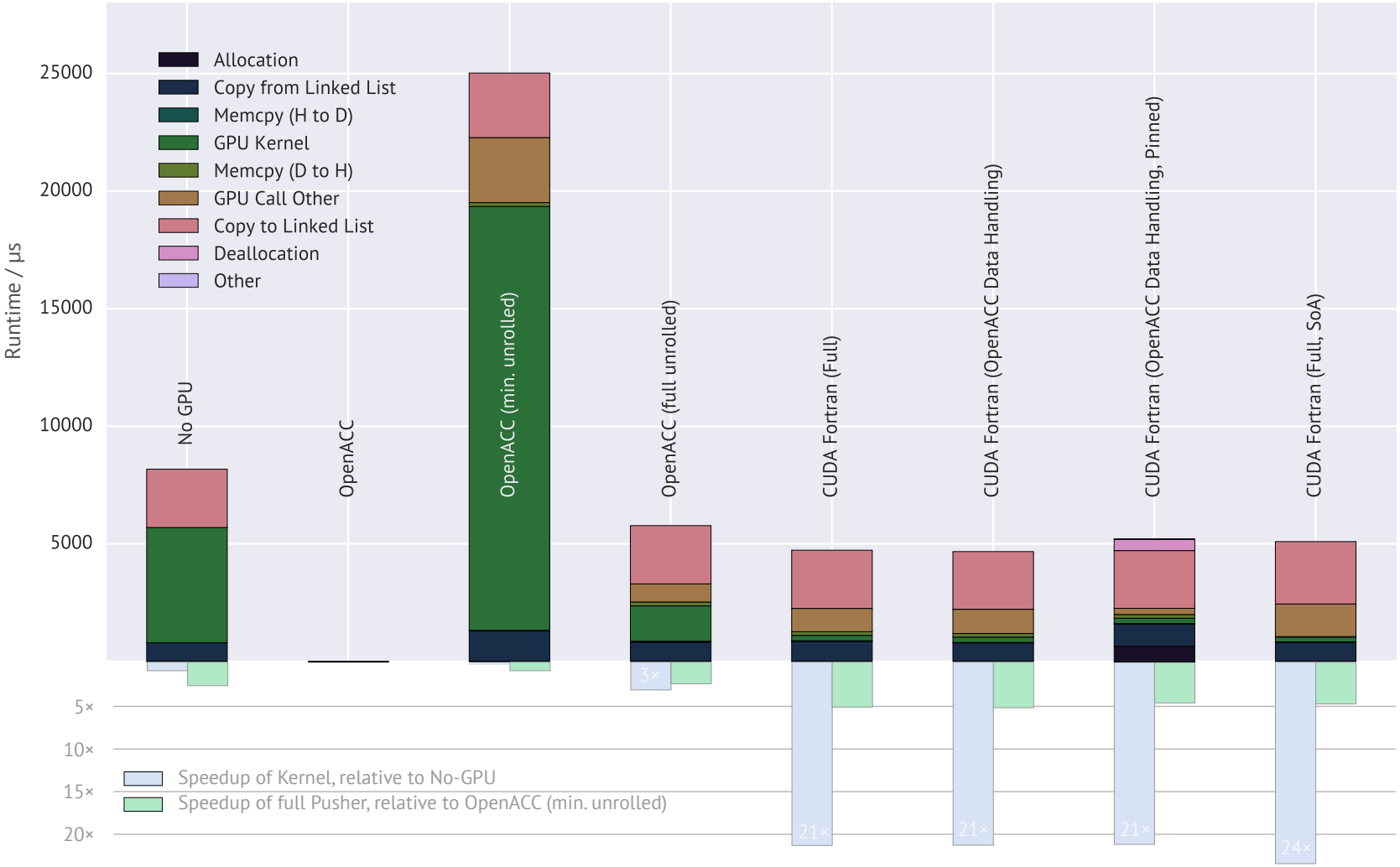
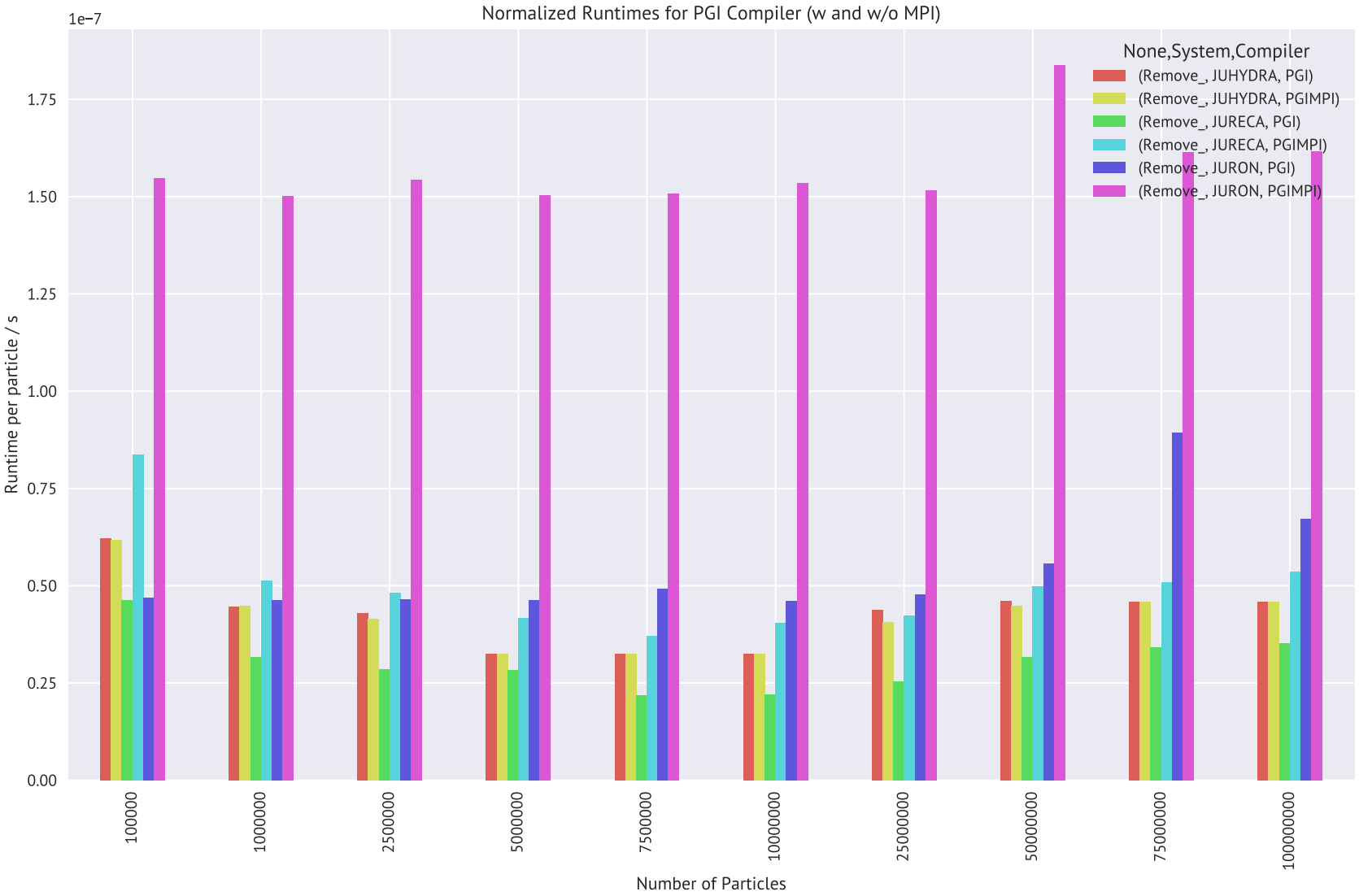


Figure: See GTC poster for details [8].

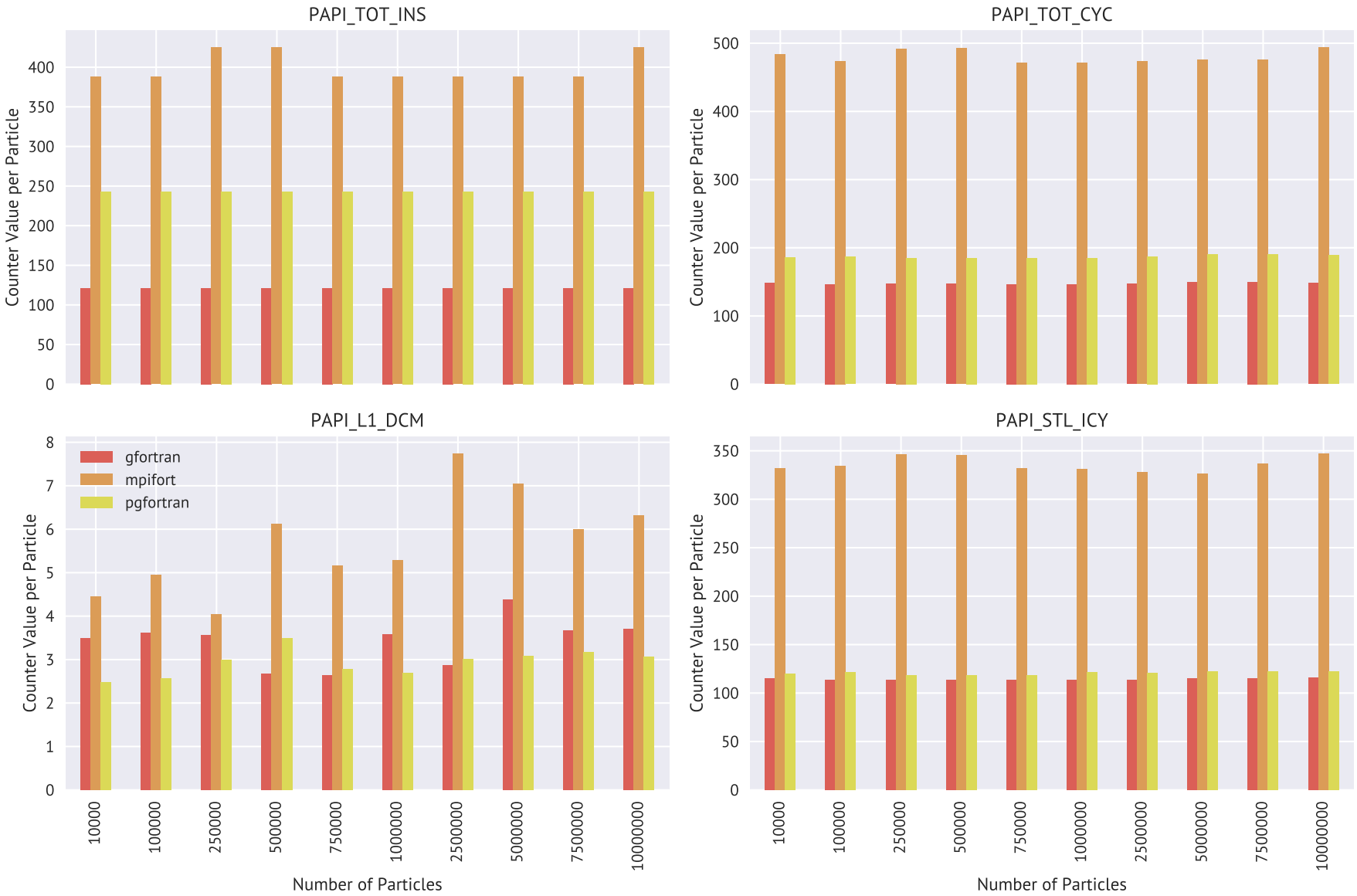
Linked List: Time for Remove on JURON

For different compilers



Selected Performance Counters on JURON

For different compilers



References: Images, Graphics I

[1] Forschungszentrum Jülich. *Hightech made in 1960: A view into the control room of DIDO*. URL: http://historie.fz-juelich.de/60jahre/DE/Geschichte/1956-1960/Dekade/_node.html (page 3).

[2] Forschungszentrum Jülich. *Forschungszentrum Bird's Eye*. (Page 3).

References I

[3] Phil Mucci and The ICL Team. *PAPI, the Performance Application Programming Interface*. URL: http://icl.utk.edu/papi/ (visited on 04/30/2017) (pages 63–66, 100).

[4] K. Germaschewski et al. "The Plasma Simulation Code: A modern particle-in-cell code with load-balancing and GPU support". In: *ArXiv e-prints* (Oct. 2013). arXiv: 1310.7866 [physics.plasm-ph] (page 90).

[5] M. Bussmann et al. "Radiative signature of the relativistic Kelvin-Helmholtz Instability". In: *2013 SC - International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*. Nov. 2013, pp. 1–12. DOI: 10.1145/2503210.2504564 (page 90).

References II

[6] Paul F. Baumeister et al. “Addressing Materials Science Challenges Using GPU-accelerated POWER8 Nodes”. In: *Euro-Par 2016: Parallel Processing: 22nd International Conference on Parallel and Distributed Computing, Grenoble, France, August 24-26, 2016, Proceedings*. Ed. by Pierre-François Dutot and Denis Trystram. Cham: Springer International Publishing, 2016, pp. 77–89. ISBN: 978-3-319-43659-3. DOI: 10.1007/978-3-319-43659-3_6. URL: http://dx.doi.org/10.1007/978-3-319-43659-3_6 (page 90).

References III

[7] P. F. Baumeister et al. “A Performance Model for GPU-Accelerated FDTD Applications”. In: *2015 IEEE 22nd International Conference on High Performance Computing (HiPC)*. Dec. 2015, pp. 185–193. DOI: 10.1109/HiPC.2015.24 (page 90).

[8] Andreas Herten, Dirk Pleiter, and Dirk Brömmel. *Accelerating Plasma Physics with GPUs (Poster)*. Tech. rep. GPU Technology Conference, 2017 (page 91).

[9] Philip J. Mucci et al. “PAPI: A Portable Interface to Hardware Performance Counters”. In: *In Proceedings of the Department of Defense HPCMP Users Group Conference*. 1999, pp. 7–10 (page 100).

Glossary I

CUDA Computing platform for GPUs from NVIDIA. Provides, among others, CUDA C/C++. 2, 22, 23, 24, 26, 27, 28, 29, 30, 31, 32, 33, 34, 86, 87

FZJ Forschungszentrum Jülich, a research center in the west of Germany. 3, 98

JSC Jülich Supercomputing Centre operates a number of large and small supercomputers and connected infrastructure at FZJ. 3

JuSPIC Jülich Scalable Particle-in-Cell Code. 2, 9, 10, 11, 26, 27, 28, 29, 30, 31, 63, 64, 65, 66, 75, 76, 77, 86, 87, 90

Glossary II

MPI The Message Passing Interface, a communication message-passing application programmer interface. 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 86, 87

NVIDIA US technology company creating GPUs. 3, 4, 5, 6, 7, 8, 89, 98

NVLink NVIDIA's communication protocol connecting CPU $\leftrightarrow$ GPU and GPU $\leftrightarrow$ GPU with 80 GB/s. PCI-Express: 16 GB/s. 4, 5, 6, 7, 8, 98

OpenACC Directive-based programming, primarily for many-core machines. 2, 14, 15, 16, 17, 18, 21, 22, 23, 24, 32, 33, 34, 37, 38, 39, 40, 41, 86, 87, 88, 91

Glossary III

P100 A large GPU with the Pascal architecture from NVIDIA. It employs NVLink as its interconnect and has fast *HBM2* memory. 2, 4, 5, 6, 7, 8, 79, 80, 85, 86, 87

PAPI The Performance API, a interface for accessing performance counters, also with aliased names cross-platform [3, 9]. 63, 64, 65, 66

Pascal The latest available GPU architecture from NVIDIA. 98

PGI Formerly *The Portland Group, Inc.*; since 2013 part of NVIDIA. 2, 26, 27, 28, 29, 30, 31, 67, 68, 69, 70, 71, 72, 73, 86, 87

PiC Particle in Cell; a method applied in a group of (plasma) physics simulations to solve partial differential equations. 2, 10, 11, 90

Glossary IV

POWER Series of microprocessors from IBM. 2, 3, 67, 68, 69, 70, 71, 72, 73, 86, 87, 89

Tesla The GPU product line for general purpose computing computing of NVIDIA. 4, 5, 6, 7, 8