ICON NWP on GPUs

Marek Jacob[1]
ICON GPU Community

Dmitry Alexeev[2], Remo Dietlicher[3], Victoria Cherkas[3], Elsa Germann[3], Fabian Gessler[3], Daniel Hupp[4], Andreas Jocksch[5], Xavier Lapillonne[4], Christoph Müller[4], Carlos Osuna[4], Daniel Reinert[1], William Sawyer[5], Ulrich Schättler[1], Günther Zängl[1] and others

1: DWD, 2: NVIDIA, 3: C2SM, 4: MeteoSwiss, 5: CSCS

ICON:
Icosahedral
Nonhydrostatic
Model

Supercomputer by NEC

- Vector machine
  NEC SX-Aurora Tsubasa
- 130th fastest civil computer

TOP 500 The List.



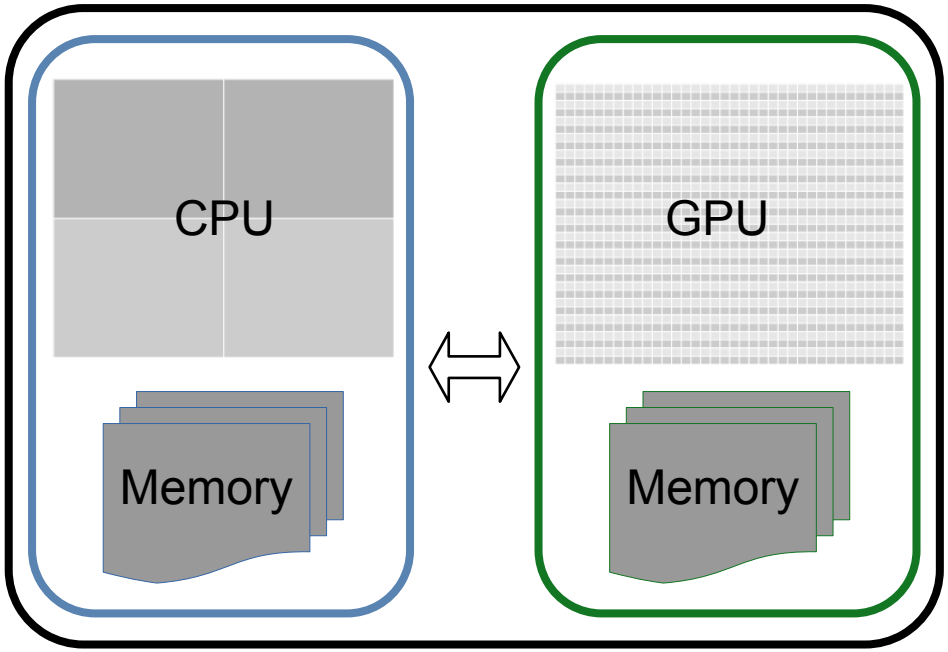
Green 500 Rank	Architecture
1 - 63	**GPU (52 x)** Other (11 x)
64	NEC Vector Computer

What is a GPU?

- Graphics Processing Unit
- Thousands of parallel processing units
- Dedicated Memory

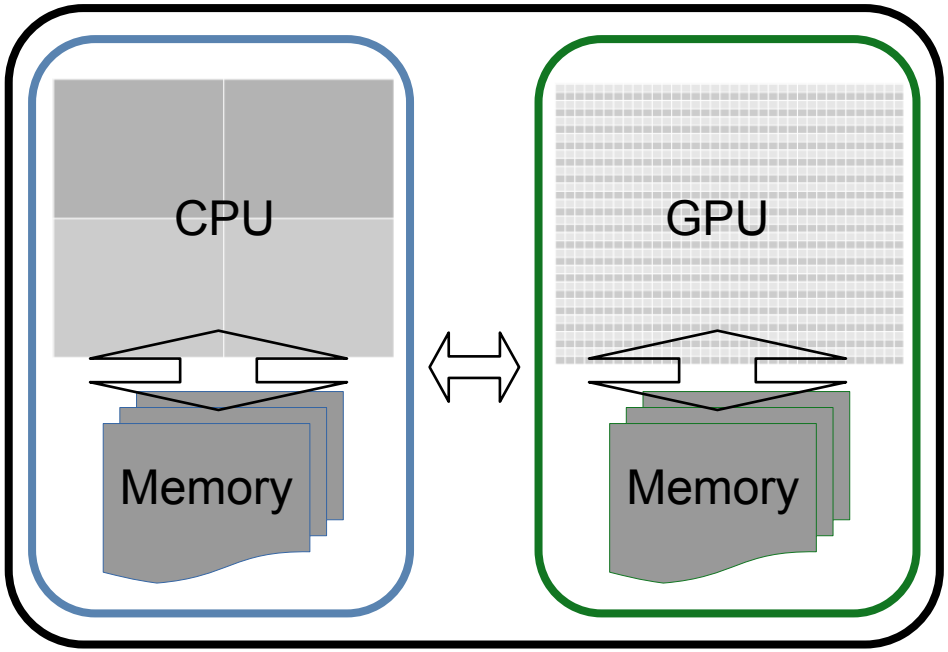
- GPU Machine: CPU + GPU
  - CPU manages GPU

➔ Special programming model required



How to program a GPU for ICON?

- ICON is written in FORTRAN
  - Use a directive based approach: OpenACC 2.6
- Thousands of parallel processing units
  - Parallelize horizontal Loops
- Dedicated Memory
  - Double Memory Management
  - Cloned Variables



OpenACC example



```
DO jt = 1, n_tracers
    ! [...] some CPU-only management code
    !$ACC PARALLEL LOOP
    DO ic = 1, n_cells
        q_tracer(ic, jt) = q_tracer(ic, jt) + ddt(ic, jt)
    END DO
END DO
```

Which version is better for GPU?

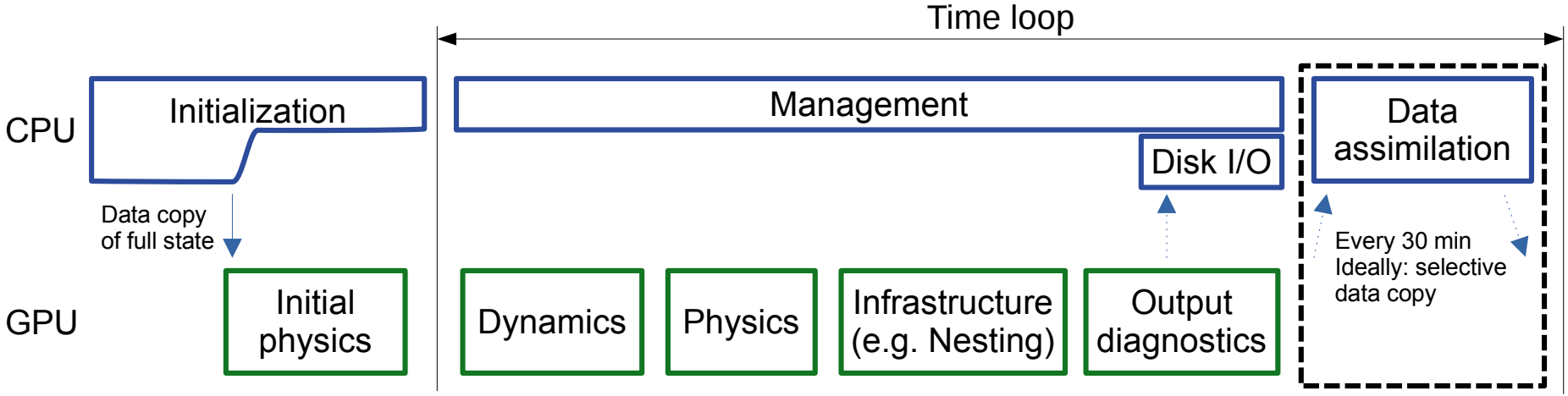

```
DO ic = 1, n_cells
    ! [...] some CPU-only management code
    !$ACC PARALLEL LOOP
    DO jt = 1, n_tracers
        q_tracer(jt, ic) = q_tracer(jt, i
    END DO
END DO
```





```
! [...] some CPU-only management code
!$ACC PARALLEL LOOP COLLAPSE(2)
DO ic = 1, n_cells
    DO jt = 1, n_tracers
        q_tracer(jt, ic) = q_tracer(jt, ic) + ddt(jt, ic)
    END DO
END DO
```

Principle program flow of ICON



- Data copy of full state to CPU is very slow
  - Data remains on GPU
  - All components in time loop should run on GPU

“ICON on GPU” == particular configurations run on GPU

Ported components:

- Dynamics
- Advection
- Radiation: RTE+RRTMGP and ecRad, radiative heating
- Microphysics (1 Moment: Graupel, Cloudice schemes)
- Turbulence (Raschendorfer, Vdiff), convection (Tiedtke-Bechtold)
- Surface (TERRA, flake, JSBACH)
- Non-orographic gravity wave drag, sub-grid scale orographic drag
- Nesting, Latent-heat-nudging, restart
- Several output diagnostics (incl. LPI)

Work-in-Progress

- 2 Moment Seifert-Beheng
- Snowpolino
- Sea-ice
- SPPT
- Pollen emission, transport, sedimentation (ICON-ART)
- EMVORADO
- ICON-Ocean

Applications of ICON on GPU

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

- Direct simulation of Quasi-Biennial oscillation in a changing climate: QUBICC project
  - Giorgetta et al. (accepted, 2022) *https://doi.org/10.5194/egusphere-2022-152*
  - Climate mode AES (formally know as ECHAM physics)

MAX-PLANCK-INSTITUT FÜR METEOROLOGIE

DKRZ DEUTSCHES KLIMARECHENZENTRUM

C2SM Center for Climate Systems Modeling

CSCS Centro Svizzero di Calcolo Scientifico Swiss National Supercomputing Centre

- Limited Area NWP
  - MeteoSwiss plans to run operationally in 2023
  - Global to Regional ICON Digital Twin: GLORI project (DWD, KIT, FZJ, MeteoSwiss, CSCS, MeteoAM, CMCC, arpae, ItaliaMeteo)
- Global NWP
  - GLORI

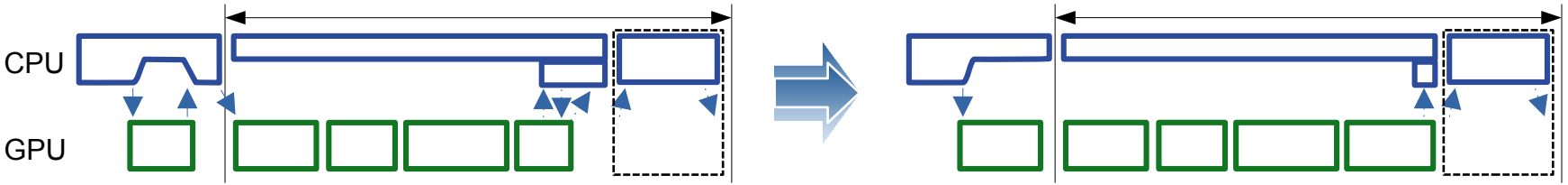
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

MeteoSwiss

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

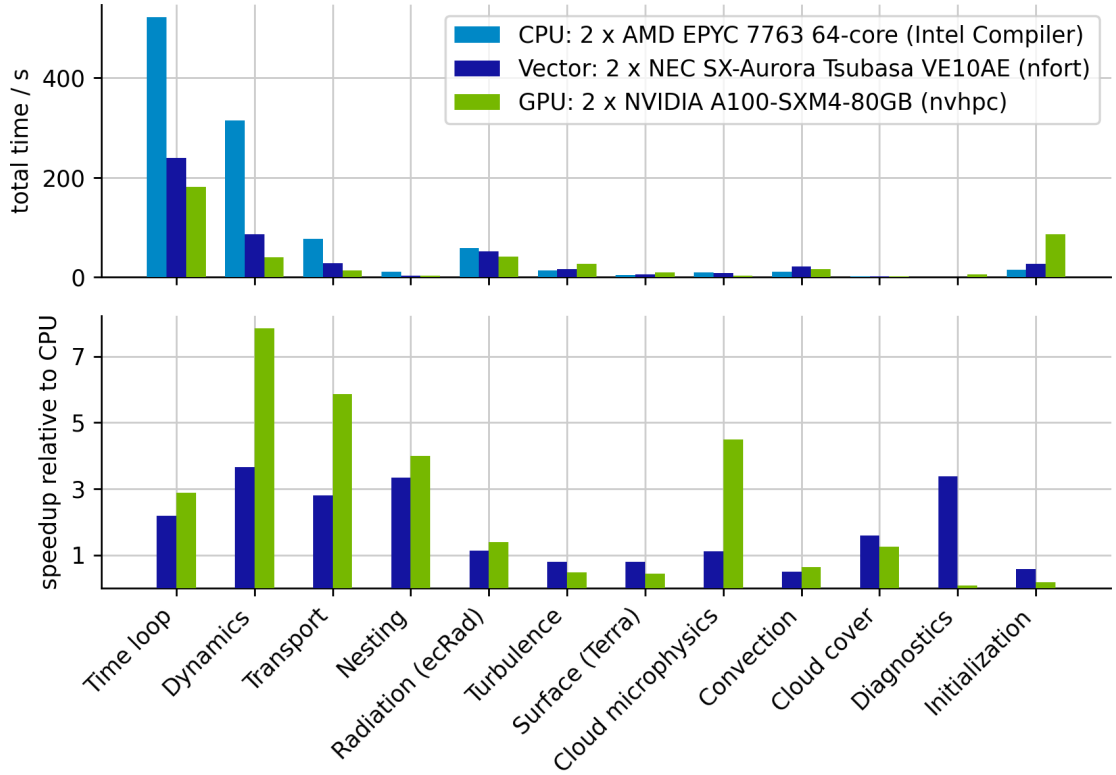
Optimizations

- GPU Code runs asynchronously to CPU → Speedup 1.16
- Fewer and fewer rare components on CPU
  - e.g. physics during initialization, diagnostics for output



- Parallel loops as long as possible
- Combine kernels `GANG(STATIC:1)`

Current performance

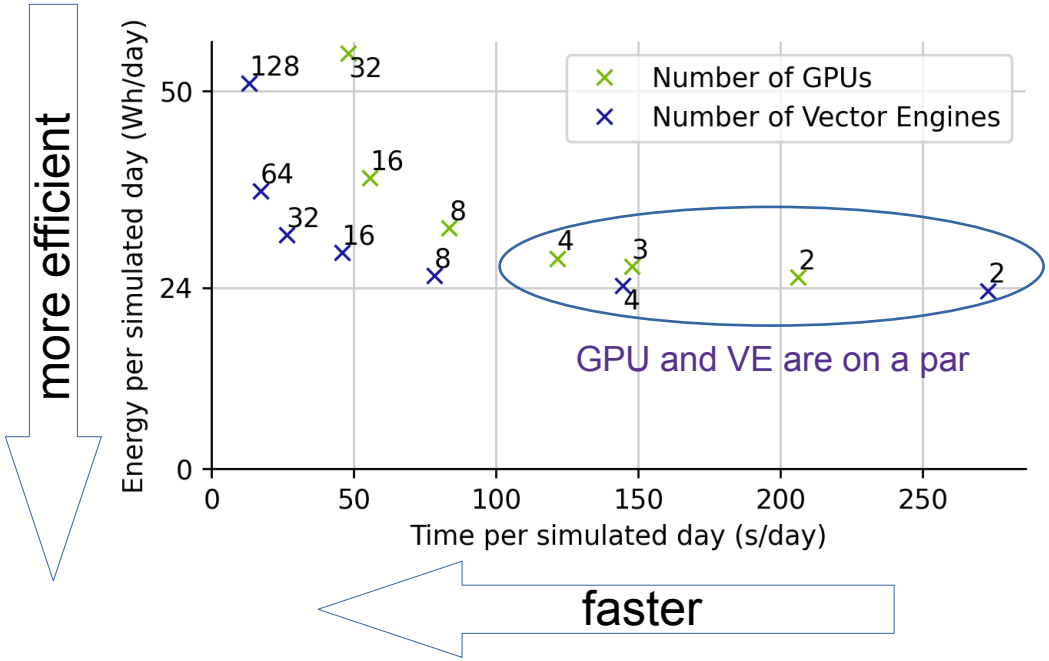


Experiment:
~ DWD ensemble forecast

- Global R2B6
  ~ 328 000 cells (40 km)
- + Nested grid over Europe
  ~ 49 000 cells (20 km)
- Output disabled

- GPU work in progress
- GPU not yet optimized for this experiment and machine

Energy efficiency



As reported by nvidia-smi (NVIDIA) and veda-smi (NEC)

Time loop only

Excluding the CPU host

GPU: out-of-the-box performance
Vector Engine: optimized setup

- GPU work in progress
- GPU not yet optimized for this experiment and machine

Final remarks

Challenges:

- Several institutions improve the scientific quality of ICON continuously
- ICON targets different platforms at once (CPU, Vector Machines, GPU)

Solutions:

- Automatic testing (Integration tests)
- Manual Review progress in Gitlab before merging
- ICON OpenACC Style Guide
- Plan internal training course “OpenACC in ICON”

Conclusions

Challenges:

- Several institutions improve the scientific quality of ICON continuously
- ICON targets different platforms at once (CPU, Vector Machines, GPU)

Solutions:

- Automatic testing (Integration tests)
- Manual Review progress in Gitlab before merging
- ICON OpenACC Style Guide
- Plan internal training course “OpenACC in ICON”

- GPUs are more and more common in super computers
- GPUs are energy efficient

- ICON has been ported to GPU using OpenACC
- Same code base for CPU, GPU and vector machines

- Close to operational deployment in regional mode

- Speedups
  - CPU socket vs. GPU: 3x and more
  - components vary, up to 8x
  - Not as good as GPU COSMO yet

- Ongoing optimization efforts on target GPU machines

Thank you for your attention!

- DKRZ is acknowledged for compute hours on Levante (CPU + GPU)
- Colleges at MeteoSwiss, C2SM, CSCS are acknowledged for pioneering and porting vast parts of ICON on GPUs
- Holger Berger (NEC) it thanked for the supporting the power measurements

Backup Slides

NEC SX-Aurora TSUBASA

- 8 Vector Engines:
  - NEC VE Typ 10AE 8-Core
  - 1.584 GHz
  - Max 300 W
  - 304 GFLOPS (double)
  - Memory: 48 GB
  - Memory Bandwidth: 1.35 TB/s
  - Vector length: 256
- Host:
  - AMD EPYC 7008 / 5376 24-Core
  - Memory: 512 GB

DKRZ Levante GPU-Node

- 4 GPUs:
  - Nvidia A100-SXM4-80GB
  - 1.275 – 1.410 GHz
  - Max: 400 (500) W
  - 9 746 GFLOPS (double)
  - Memory: 80 GB
  - Memory Bandwidth: 2.039 TB/s
  - Cuda cores (double): 3456
- Host:
  - AMD 7713 64-Core 2.0 – 3.7 GHz
  - TDP: 225 (225 – 240) W
  - Memory: 512 GB

- Atos BullSequana XH2000

DKRZ
DEUTSCHES
KLIMARECHENZENTRUM

- CPU-only Host:
  - AMD 7763 64-Core
  - 2.45 – 3.5 GHz
  - TDP: 280 (225-280) W
  - Memory: 256 GB
  - Per Socket Mem BW: 204.8 GB/s