Artificial Intelligence for Simulation

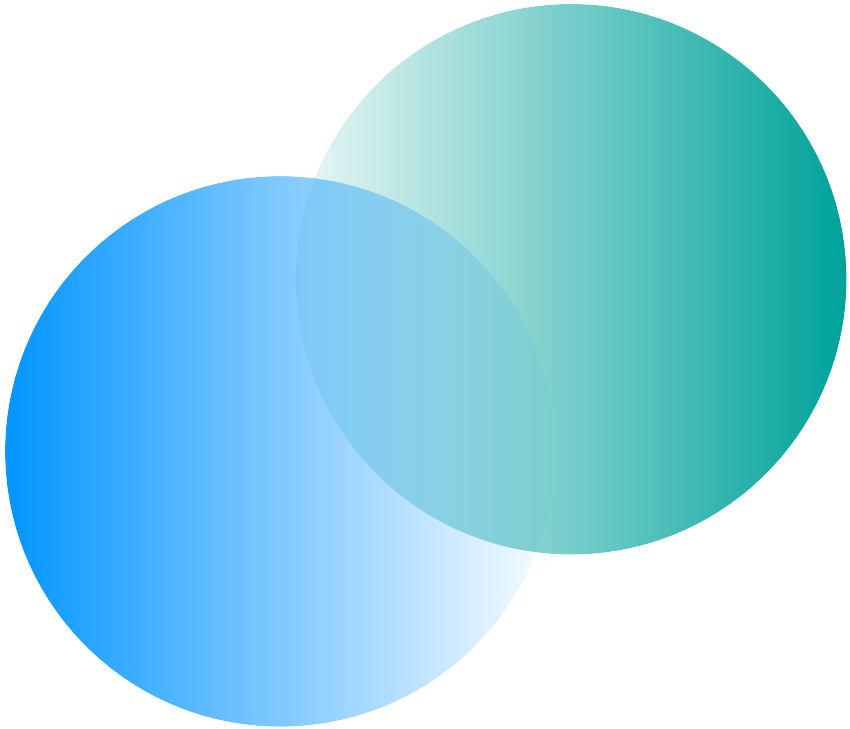
an R&D Scientific ML program

Improving the radiative scheme with machine learning on a heterogeneous cluster

Christophe Bovalo, Rémi Druilhe – *Atos BDS R&D AI4sim*
Matthew Chantry, Peter Düben - *ECMWF*

Atos

ML emulation of the radiation scheme



Radiation schemes in IFS

ecRad

What is ecRad?

- ecRad is a library aiming at simulating the radiation scheme through 5 different solvers:
  - *Monte Carlo Independent Column Approximation (McICA)*
  - *Tripleclouds*
  - *SPARTACUS*
  - *Homogeneous (plane parallel) solver*
  - *Cloudless solver*

- It computes vertical profiles of solar (shortwave) and near-infrared (longwave) fluxes and heating rates

- It is tightly coupled with the IFS but can run offline

Atos

Emulation of the radiation scheme

Context

- Radiation scheme represents less than 5% of the computational time of IFS
  - Run on a coarser grid
  - Not called every time step

- SPARTACUS is a solver that simulates the 3D radiative effects of clouds but it is too expensive to be run in IFS operational configuration

- Tripleclouds represents cloud heterogeneity via three regions at each height (Shonk and Hogan, 2008)
- SPARTACUS = Tripleclouds + 3D radiative effects

- The idea is to learn the difference between the outputs of SPARTACUS and Tripleclouds as a corrective term to the Tripleclouds formulation (rather than learning directly the entire SPARTACUS outputs)

- Additional information can be found on https://git.ecmwf.int/projects/MLFET/repos/maelstrom-radiation/browse

Atos

Data
Inputs and outputs

Inputs: same inputs as for Tripleclouds

Scalar inputs: solar irradiance, cosine of solar zenith angle, skin temperature, shortwave albedo (albedo band), direct shortwave albedo (albedo band), longwave emissivity (emissivity band)

Column inputs (on 137 levels): specific humidity, gas mixing ratio, aerosol mass mixing ratio (aerosol type), cloud fraction, liquid water mixing ratio, ice water mixing ratio, liquid effective radius, ice effective radius

Half-level inputs (on 138 half-levels): pressure, temperature

Level interface inputs (136 levels): cloud overlap parameter

Outputs: Difference between the outputs of SPARTACUS and Tripleclouds

Downward shortwave (SW) flux

Upward shortwave flux

Downward longwave (LW) flux

Upward longwave flux

Longwave heating rate

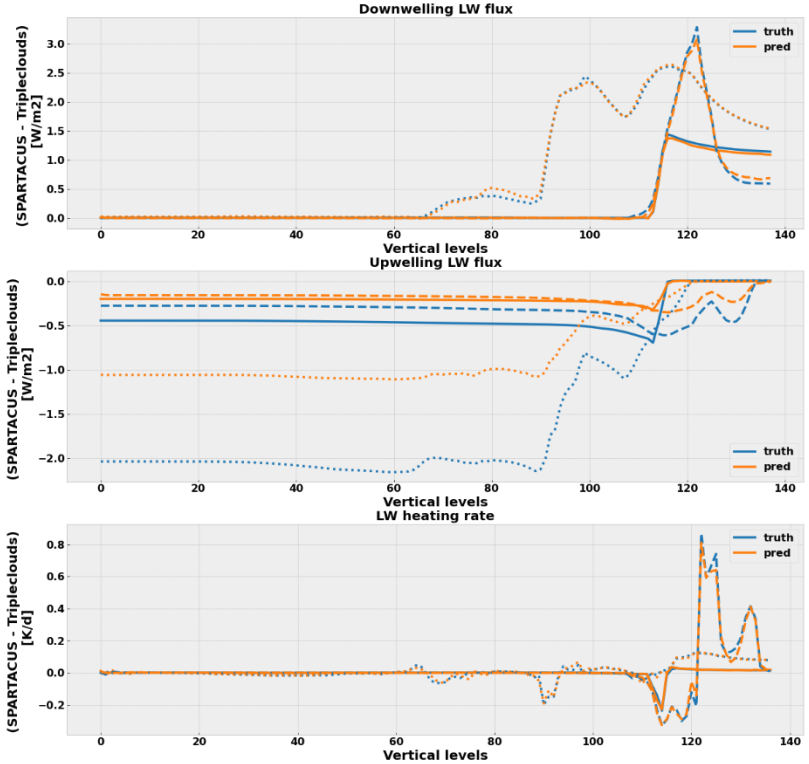
Shortwave heating rate

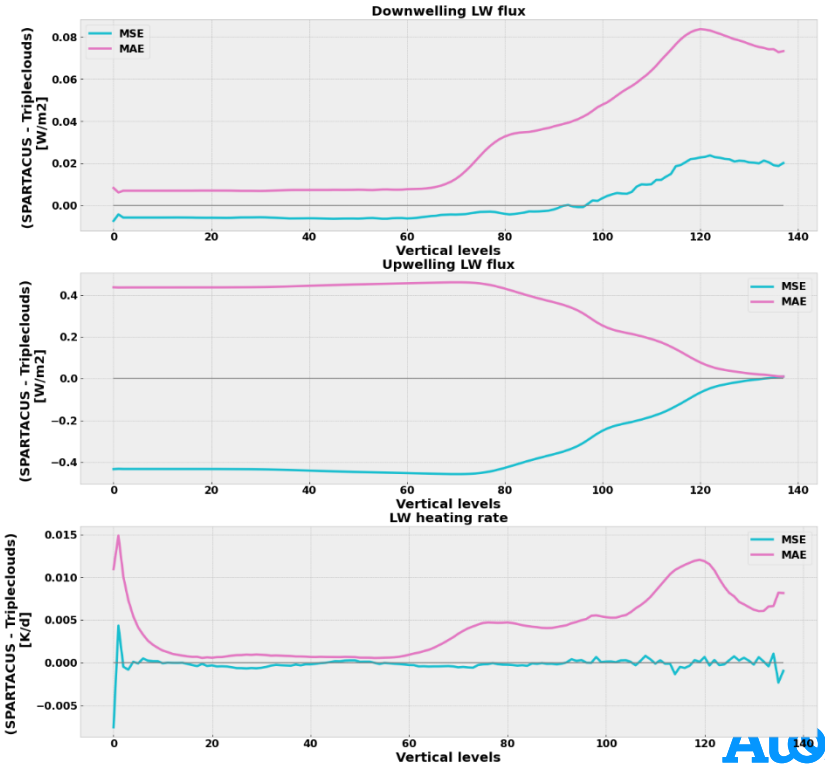
(Dataset downloaded using climetlab and the climetlab-maelstrom-radiation plugin)

Atos

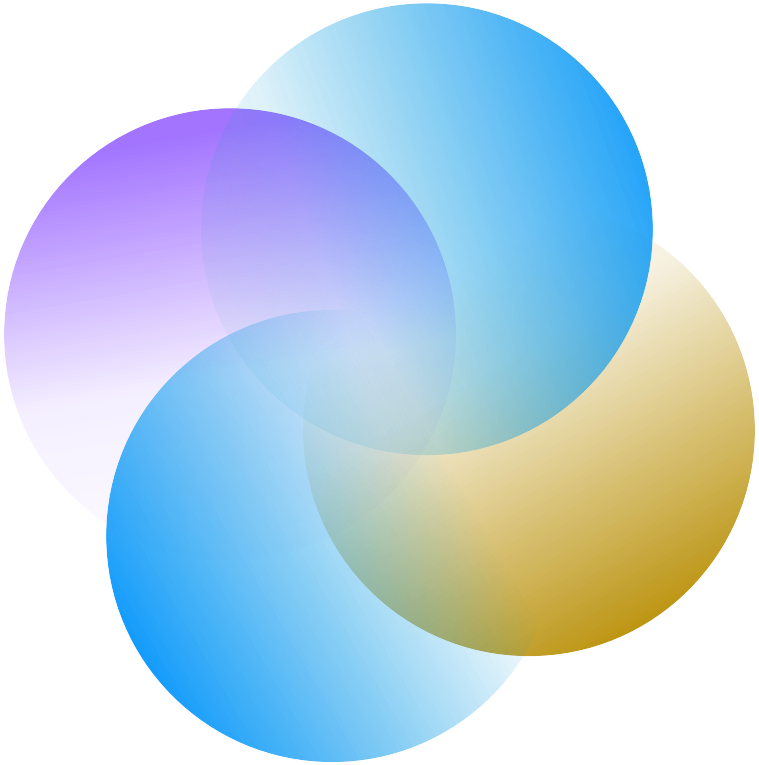
Preliminary results

- Model architectures: 1D CNN with MultiHeadAttention layers (SW and LW)
- Combined loss: loss for the fluxes (MSE) + loss for the heating rates (MSE – using a custom layer)
- Better results right now for the LW radiation but work still in progress (upward fluxes more difficult to predict)





Coupling ecRad ML to the IFS



Heterogeneity of the clusters

- In a close future, hardware heterogeneity will be a matter of fact in clusters. Standard CPU nodes will coexist with AI-accelerated nodes (GPU, IPU, TPU, ✱PU, FPGA,…), either on the same node (hybrid node), either on dedicated nodes (separation of standard nodes an AI-accelerated nodes)
- We make the following hypothesis: **n** CPU processes communicates with **m** AI-accelerated processes, where **n** > **m**
- Resulting issues are
  - Load balancing amongst all the AI-accelerated nodes to efficiently exploit all the processors
  - Possible bottlenecks on the AI-accelerated nodes, especially bursts of processing before a Barrier (all processes send data at the same time to the AI-accelerated nodes)

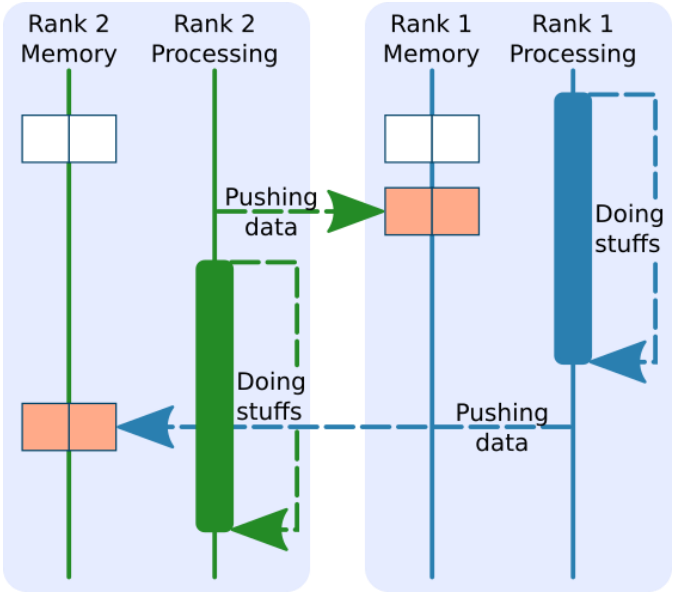
Solutions to connect a solver to an inference engine

	Pros	Cons
Zero copy	Efficient solution to transfer data between processes on the same node	Not standardized, not designed for a network of nodes
HTTP REST	Standardized, multiple implementations (some open source), easy to take in hand (lots of literature)	Designed for cloud, no direct access to memory
HPE SmartSim	Open source, working solution for coupling with ML	Not standardized, single implementation (Redis), no direct access to memory
MPI Send/Recv	Standardized, multiple implementations (some open source)	No direct access to memory
MPI RMA	Standardized, multiple implementations (some open source), direct access to memory (RDMA)	Concurrent access to memory

Investigating MPI Remote Memory Access (RMA)

- RMA has been introduced in MPI-2 (1997) and improved in MPI-3 (2012) and MPI-4 (2021)
- RMA is standardized in the HPC community and thus is included into multiple MPI implementations (Open MPI, MPICH, ...)

- Concept
  - Move data without requiring that the remote process synchronize
  - Each process exposes a part of its memory to other processes
  - Other processes can directly read from or write to this memory



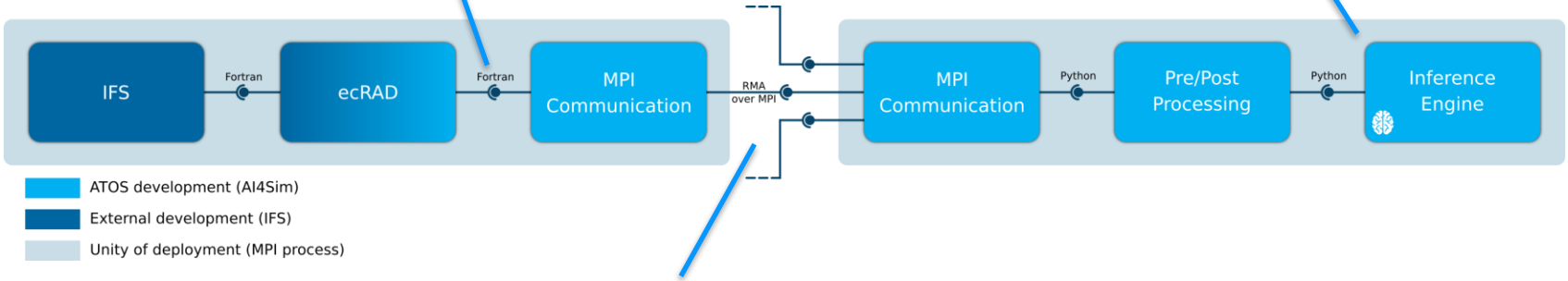
Source: Torsten Hoefler, MPI Remote Memory Access Programming (MPI3-RMA) and Advanced MPI Programming, presented at RWTH Aachen, Jan. 2019

Atos

Architecture of the connection IFS ⟷ ecRad ML

Develop a mock-up library in Fortran 90 aiming at replicating the interface between ecRad and its solvers

The inference engine is developed in Python and deployed on the NVIDIA A100 GPU nodes of our cluster



The interface communicates, using MPI RMA, in a **n** to **m** manner (**n** > **m**) with the inference engines:

- **n** CPU processes communicate with **m** GPU processes
- A "passive" load balancing is done on the GPU nodes

Atos

Diagram sequence IFS ↔ ecRad ML (synchronous)

Short terms developments

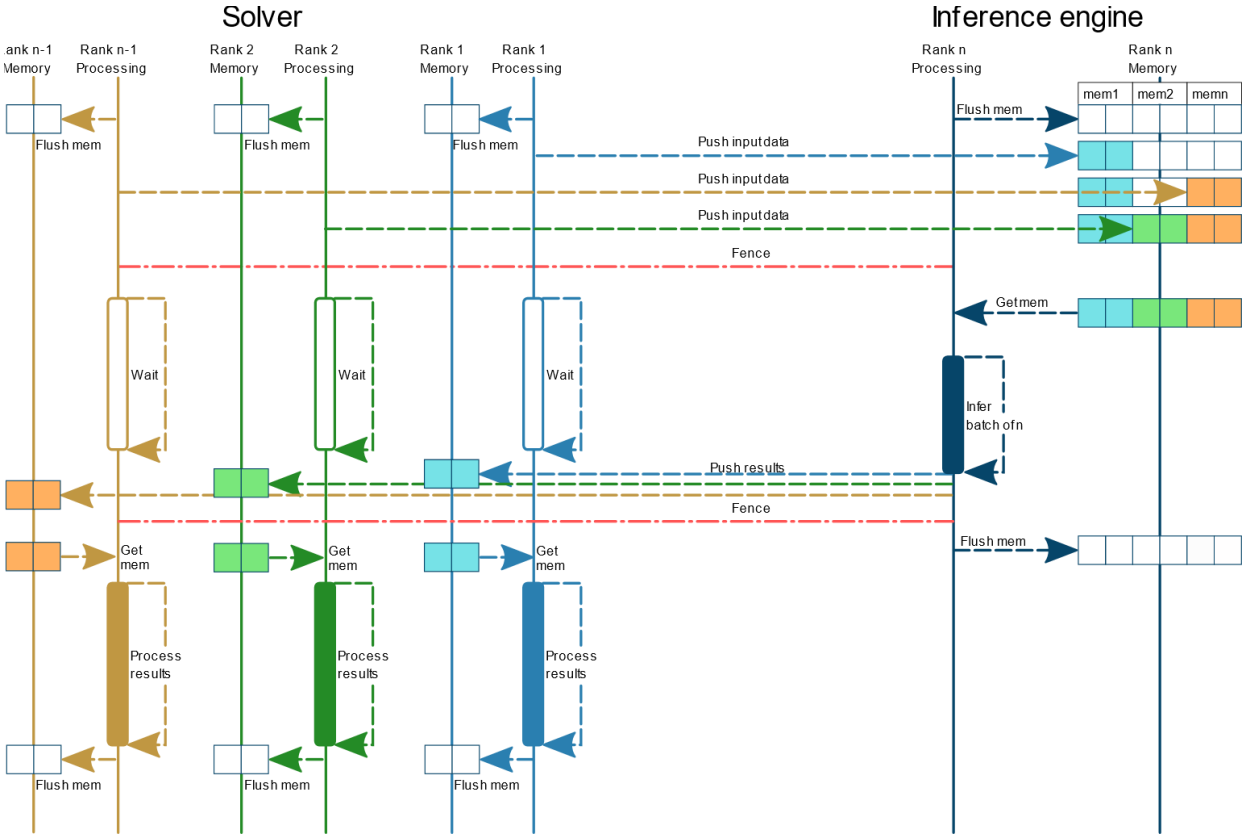
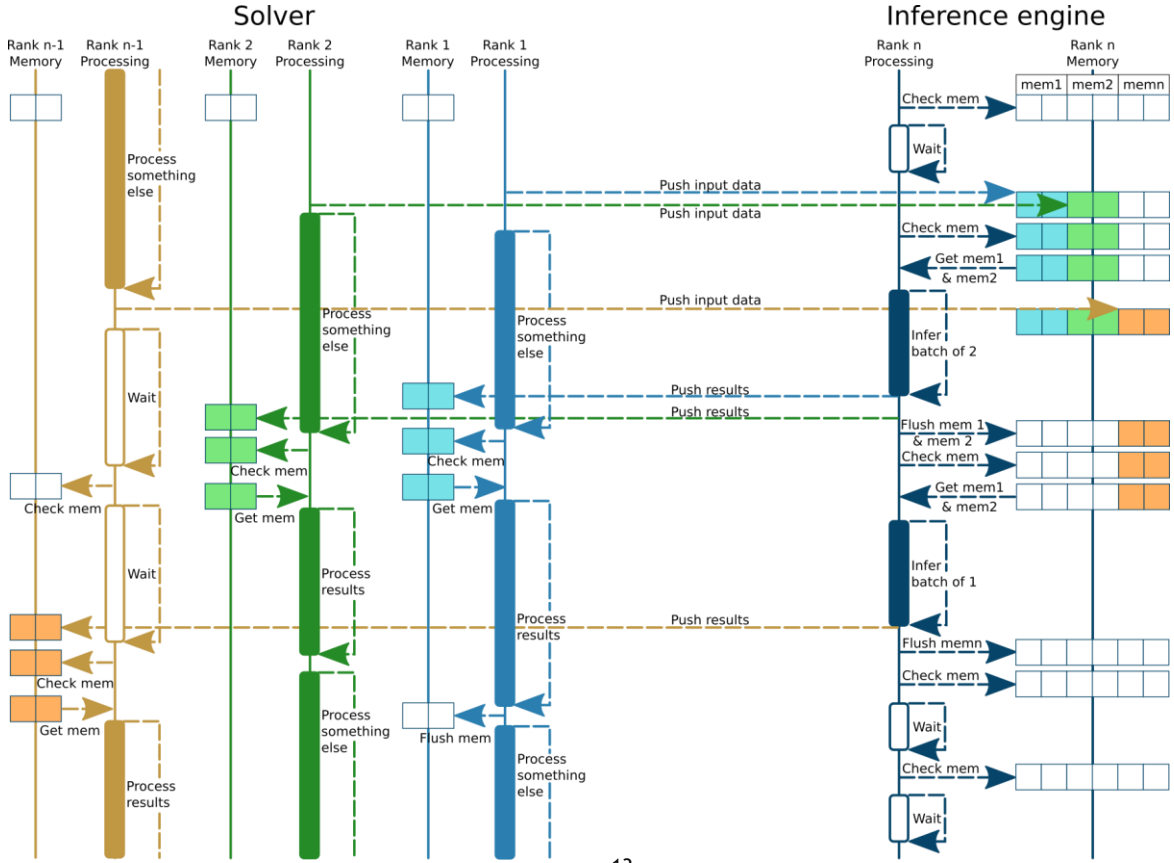


Diagram sequence IFS ↔ ecRad ML (asynchronous)

Long terms developments



Conclusion

On the model side

- SPARTACUS represents the 3D radiative effects of clouds but it is too expensive to be run in operational
- Instead of emulating the entire scheme, we try to learn a corrective term
- Preliminary results are positive for downwelling fluxes
- Improvement is therefore needed for upwelling fluxes
- Adding the outputs of Tripleclouds may help increase the accuracy of the NN

On the coupling side

- MPI RMA offers the freedom to implement synchronous and asynchronous communications patterns between the solver and the inference engine
- MPI RMA being part of the MPI standard and multiple implementations being available, vendor lock-in is not possible
- Work still in progress and results should arrive shortly

Thank you !
Do you have any questions?

For more information on AI4Sim please contact:

Product Owner
Gaël Goret
+33 683 826 720
gael.goret@atos.net

Product Manager
Matthieu Isoard
+33 651 821 763
matthieu.isoard@atos.net

Atos