Anemoi scalability improvements

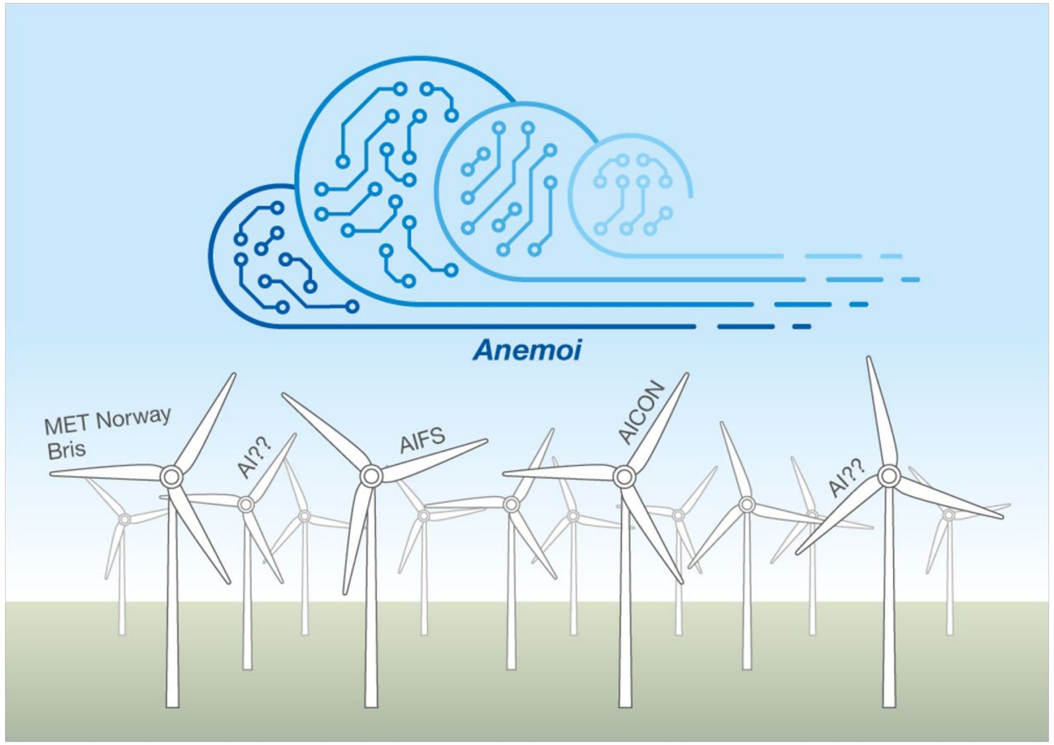
21st HPC workshop

Jan Polster and Cathal O'Brien

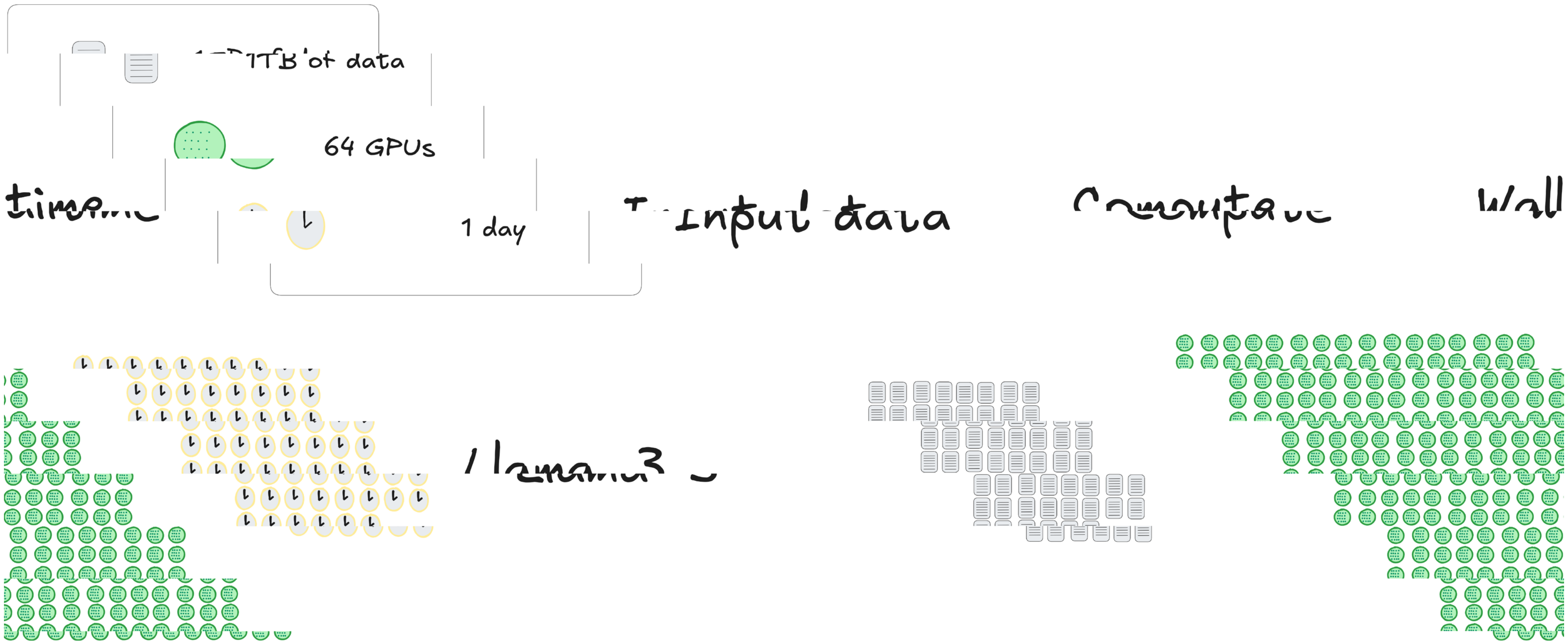
ECMWF

What is Anemoi

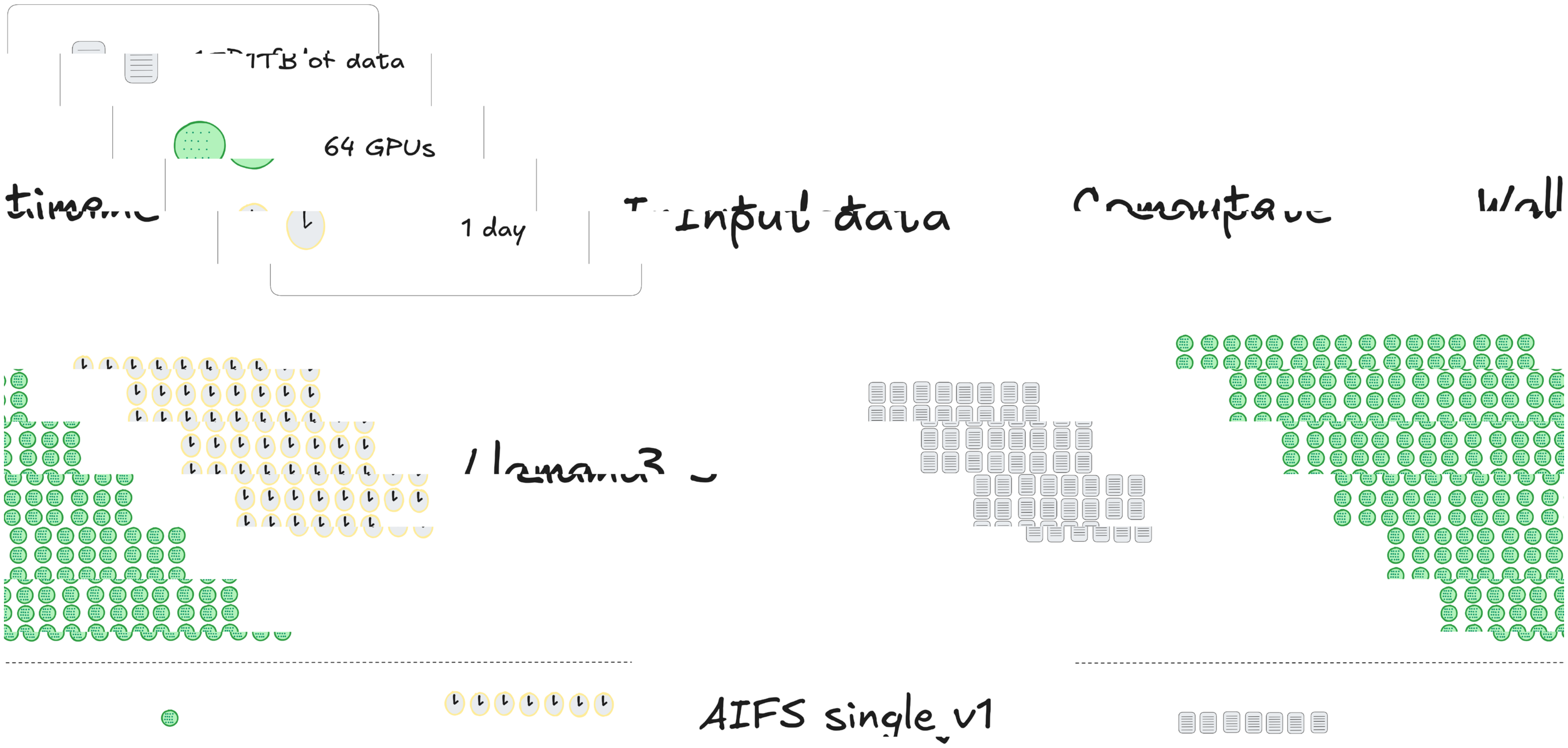
- A framework for developing data-driven weather forecasting models
- Entire framework from datasets to inference
- Multiple use-cases
  - Global, LAM, stretched grid, ensembles



2024: A tale of two models

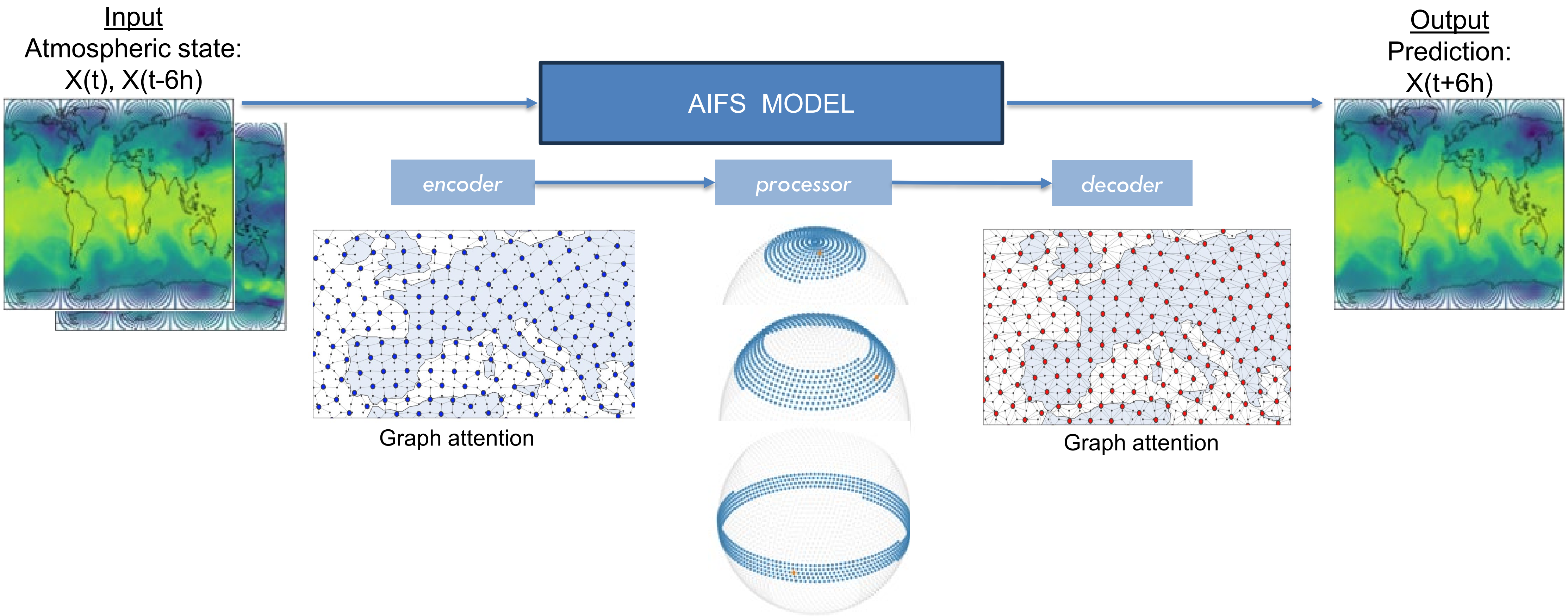


2024: A tale of two models



AIFS: Data-driven Weather Forecasting Systems

OPERATIONAL
AIFS SINGLE



Input
Atmospheric state:
X(t), X(t-6h)

AIFS MODEL

encoder

processor

decoder

Output
Prediction:
X(t+6h)

Graph attention

Graph attention

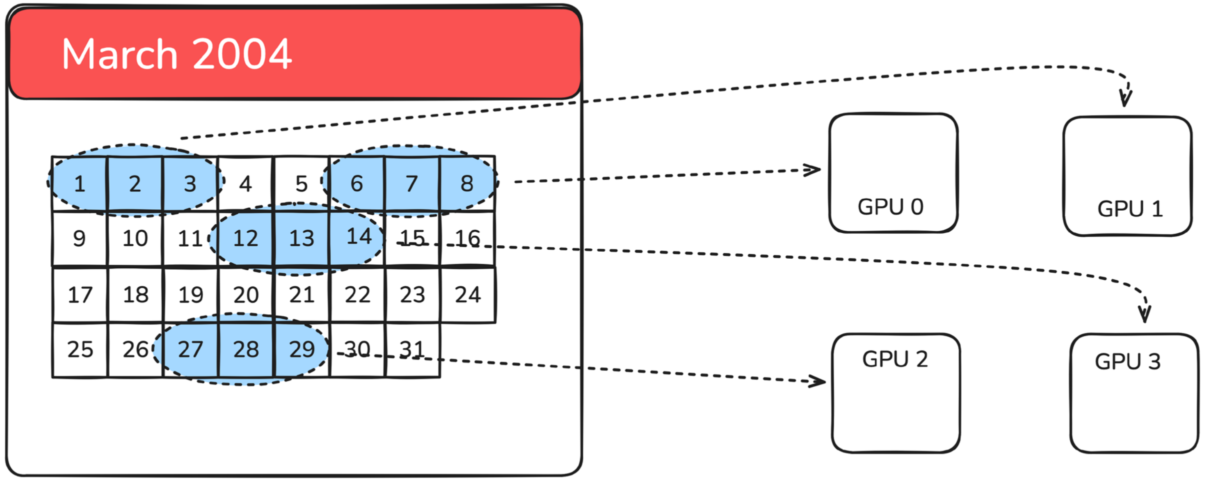
Transformer blocks and windowed attention (attention across regional bands). On a coarser grid than input grid

Implemented in

anemoi

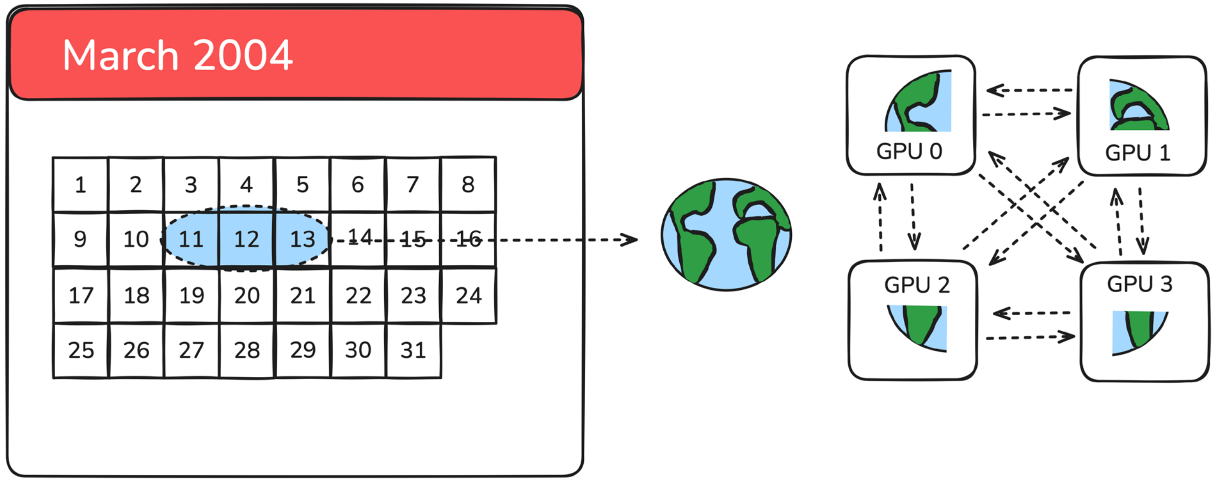
Parallelism in Anemoi

- Data Parallelism: DistributedDataParallel (PyTorch)
  - Distribute training batch across model replicas
  - Aggregate gradients via all-reduce, good scaling :)
  - Limited by batch size :(



1	2	3	4	5	6	7	8
9	10	11	12	13	14	15	16
17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	

- Model parallelism: domain-specific sharding
  - Distribute input data and activations across GPUs
  - Collective communication to handle synchronisation
  - Limited by communication overheads

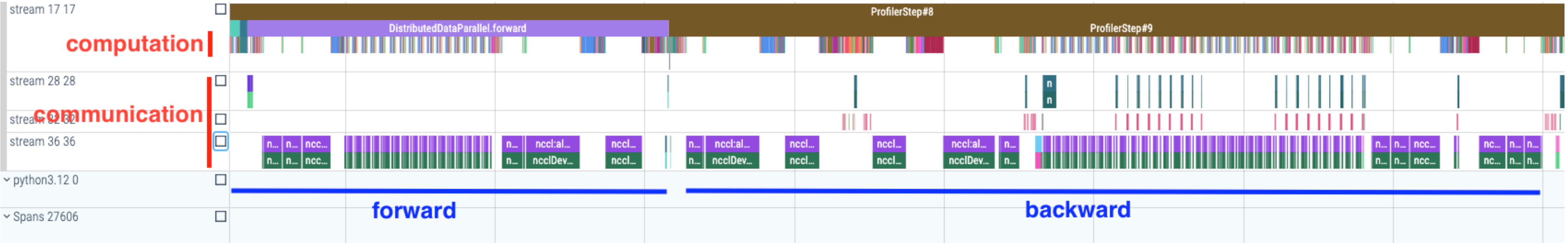


1	2	3	4	5	6	7	8
9	10	11	12	13	14	15	16
17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	

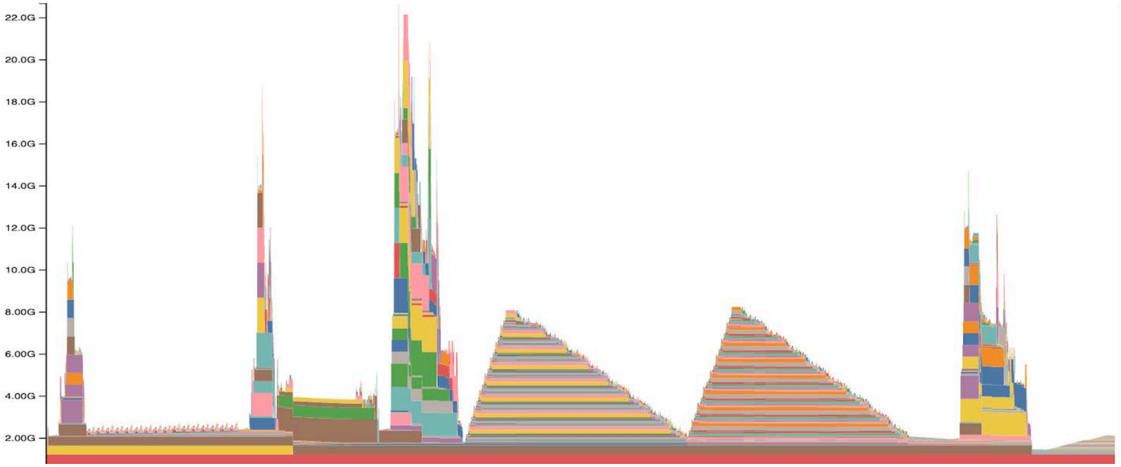
Scaling the input resolution

- **Bottlenecks**:
  1. Loading/storing full training batch on device
  2. Encoder/Decoder: communication and memory overhead

Resolution	Batch Size
O96 (100km)	0.05 GB
N320 (32km)	0.65 GB
O1280 (9km)	7.40 GB
O2560 (4km)	~30 GB



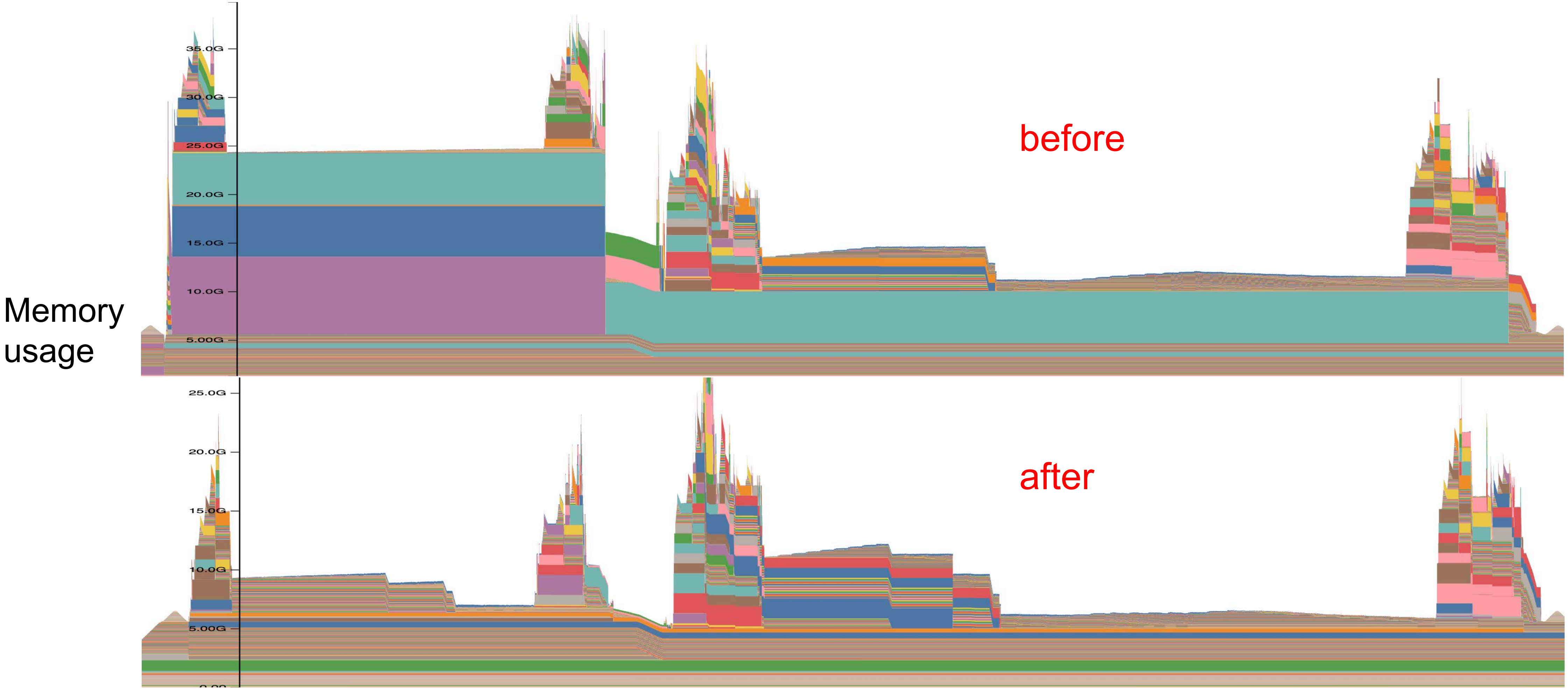
- **Improvements:**
  1. Keep batch sharded
  2.1. Mapper edge sharding
  2.2. Mapper chunking



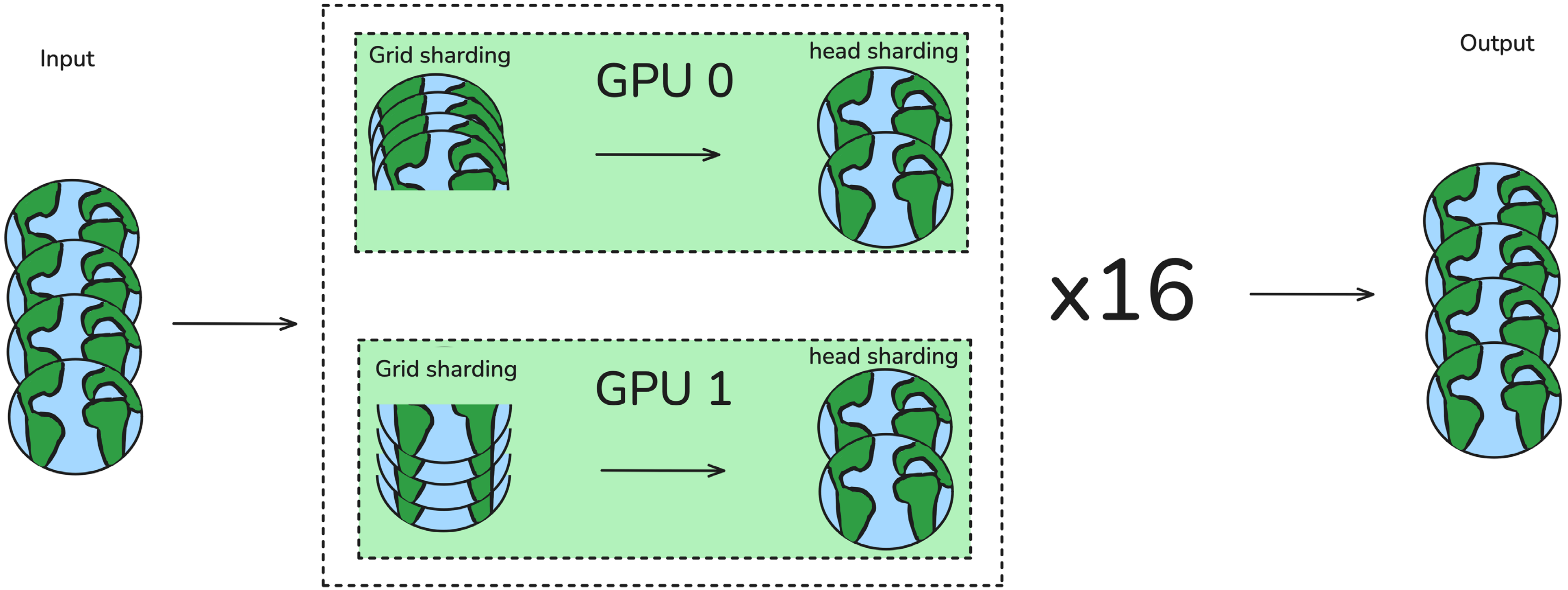
Keep batch sharded

- Avoid materialising full input/output grid in memory
    - Load batches in shards
    - Compute loss locally + all-reduce for global loss

Resolution	Batch Size
O96 (100km)	0.05 GB
N320 (32km)	0.65 GB
O1280 (9km)	7.40 GB
O2560 (4km)	~30 GB

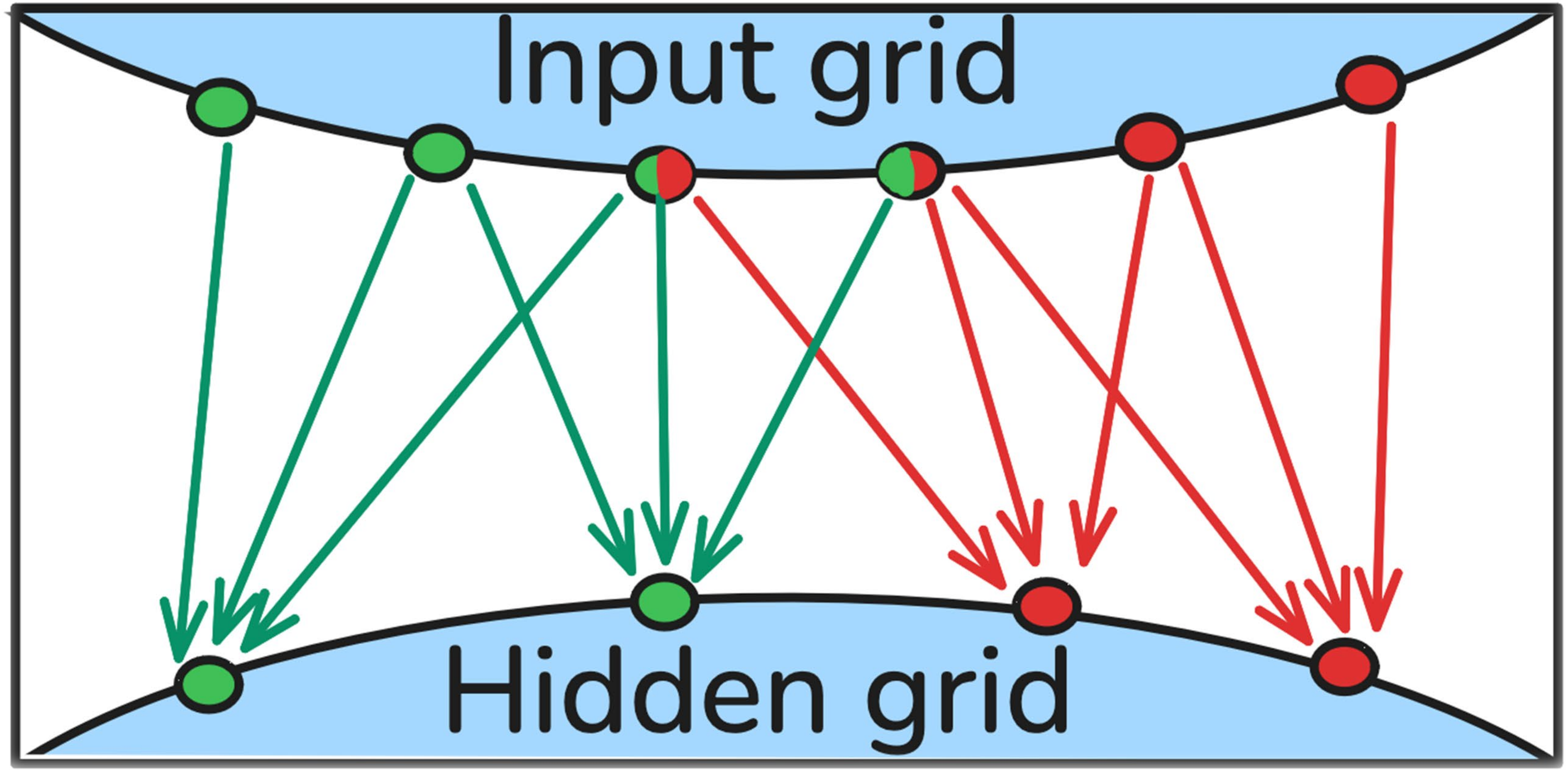


Sharding Attention



- 4x all-to-all communication per fwd/bwd
- But: we can do better by exploiting sparsity + locality of graph attention

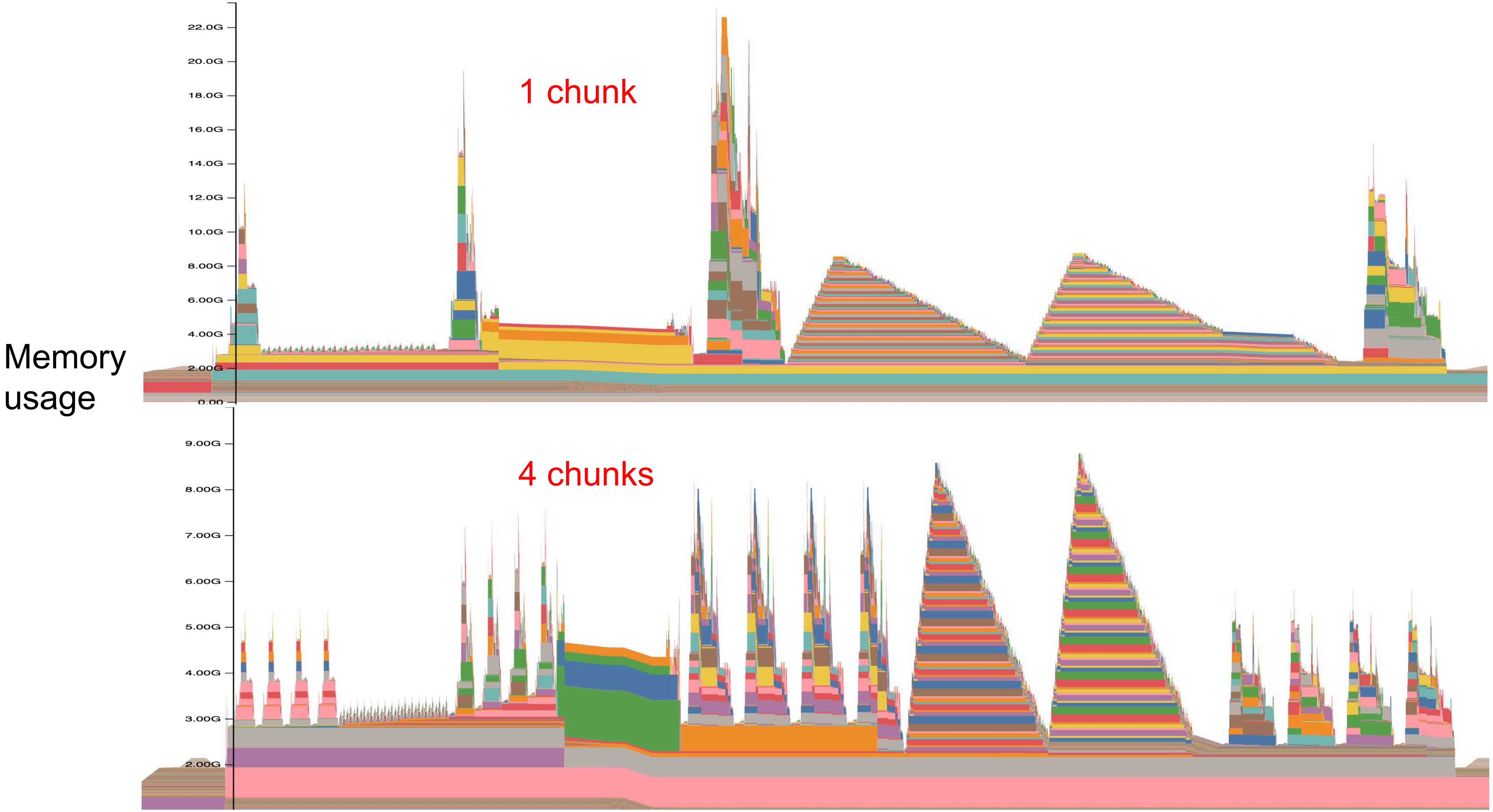
Mapper edge sharding



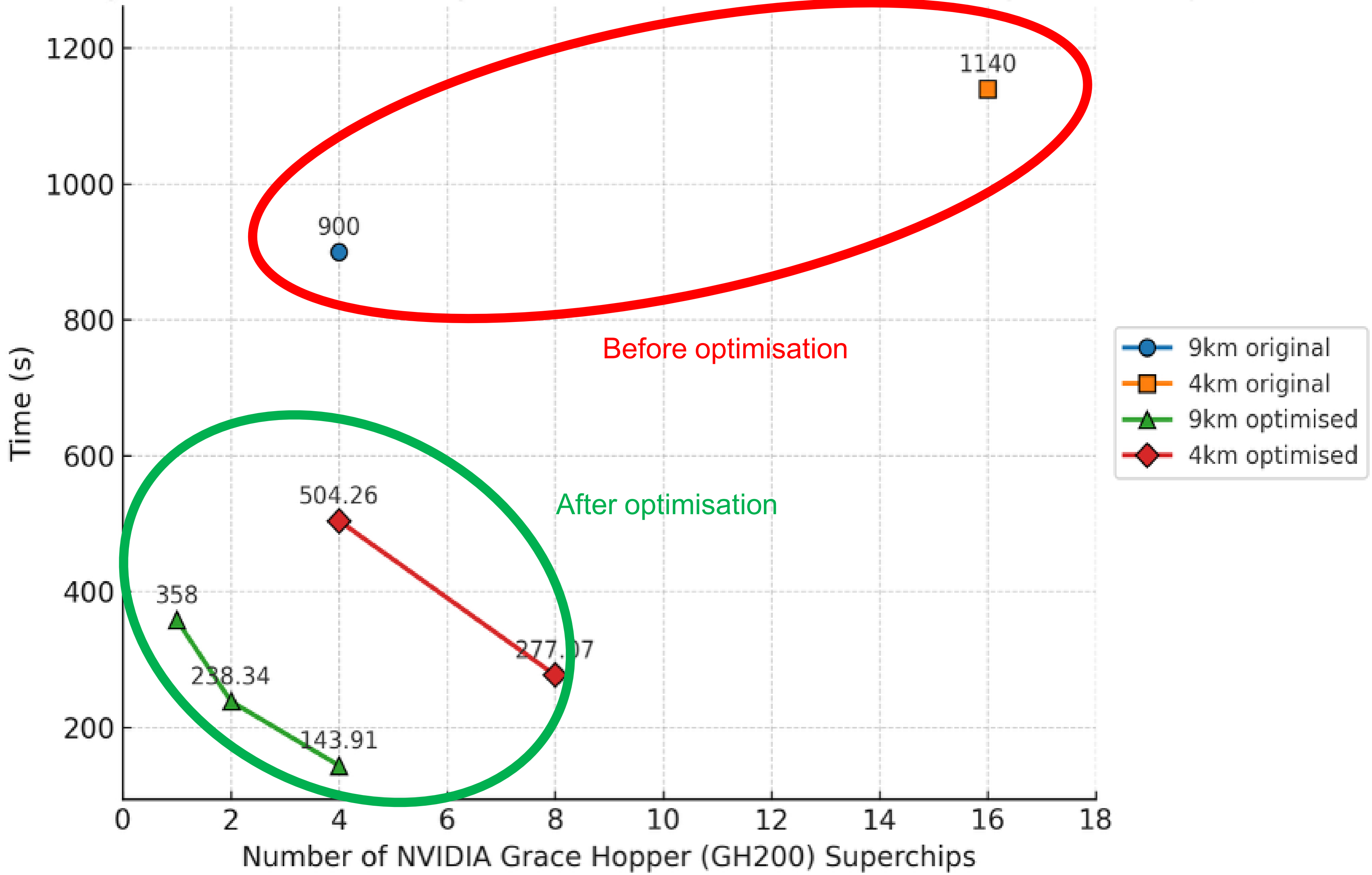
- 1x all-gather in fwd + reduce-scatter in bwd
- + independent subgraphs on each device (with small overlap)

Mapper chunking

- Same idea as edge sharding but sequentially



AIFS single inference 15-day forecast (100 fields, 150 steps, no output)



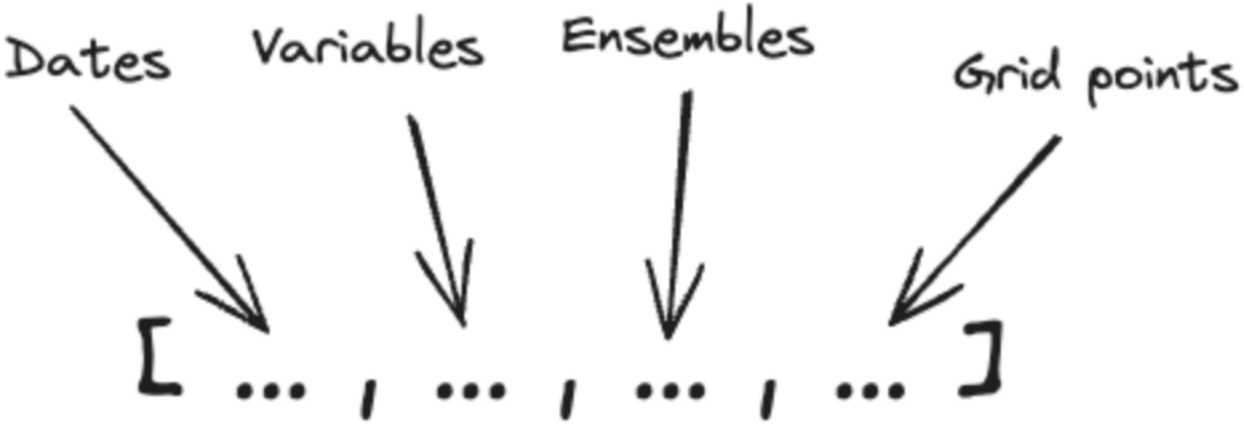
Before optimisation

After optimisation

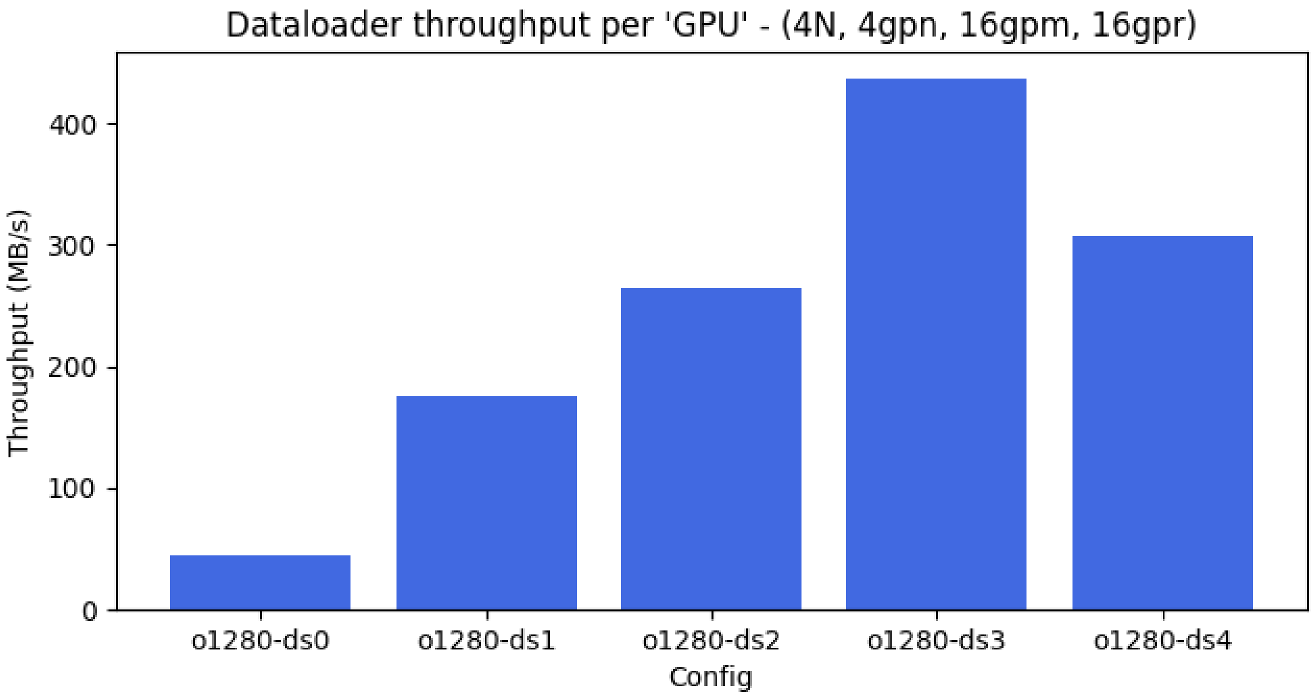
Anemoi datasets

Resolution	files per date
O96 (100km)	1
N320 (32km)	1
O1280 (9km)	4
O2560 (4km)	16

- Anemoi-datasets built on top of Zarr
  - Offers an array-like view of a collection of files
  - Each date is at least 1 file
  - Larger resolutions are split across multiple files

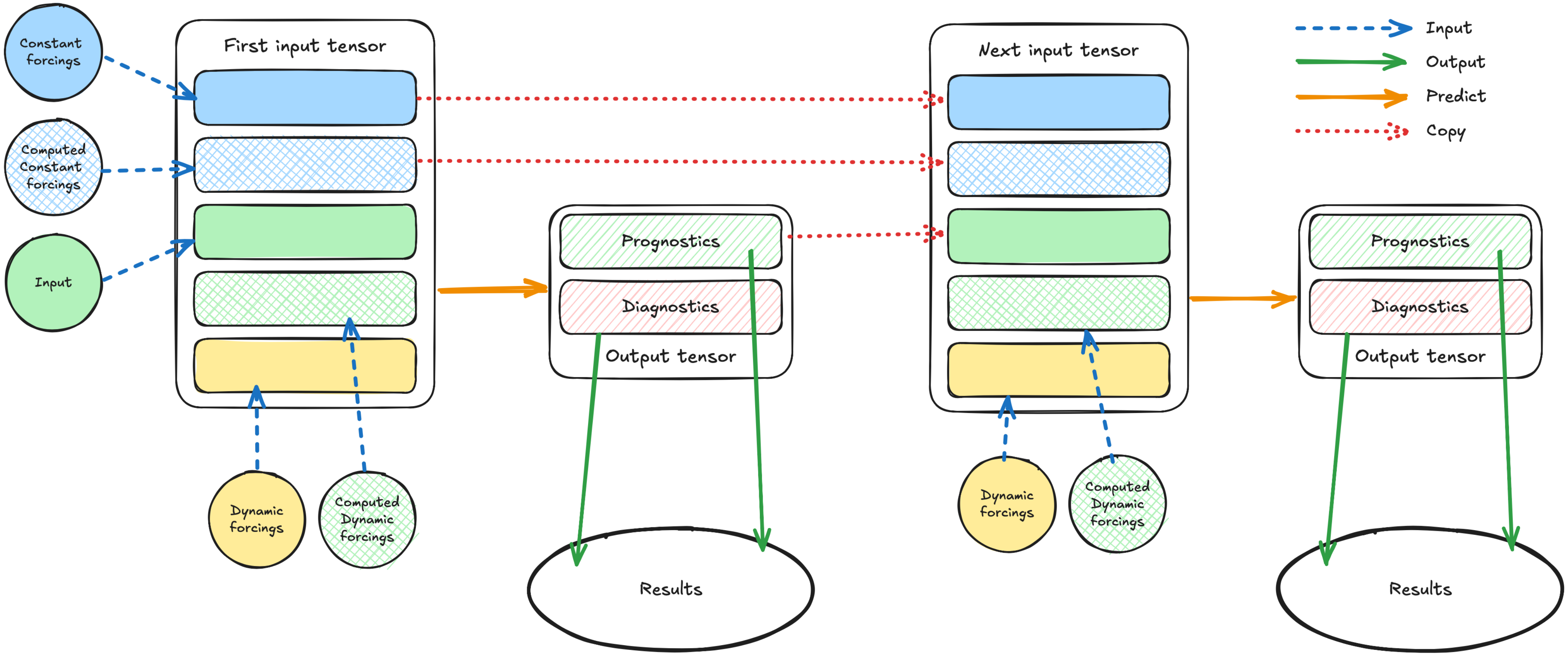


Reformatted datasets

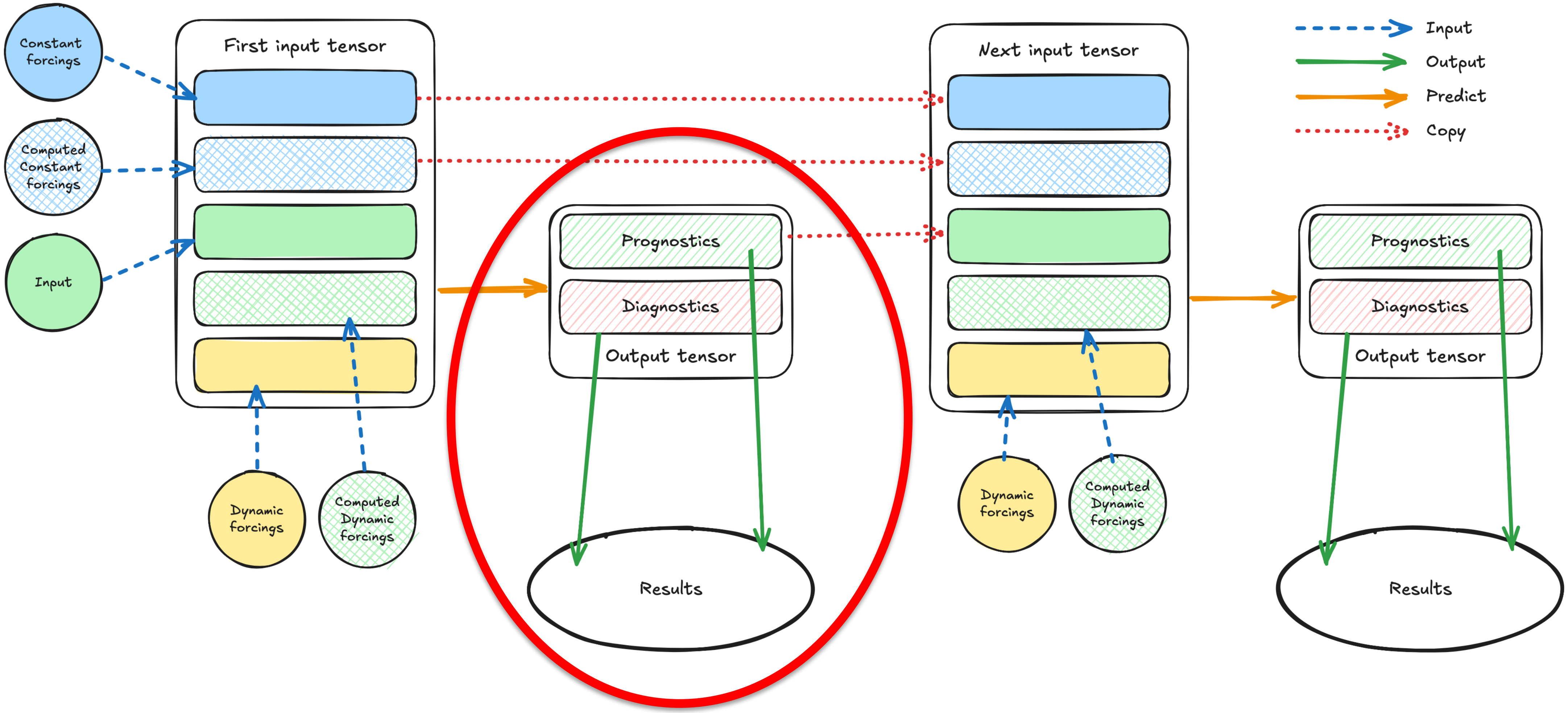


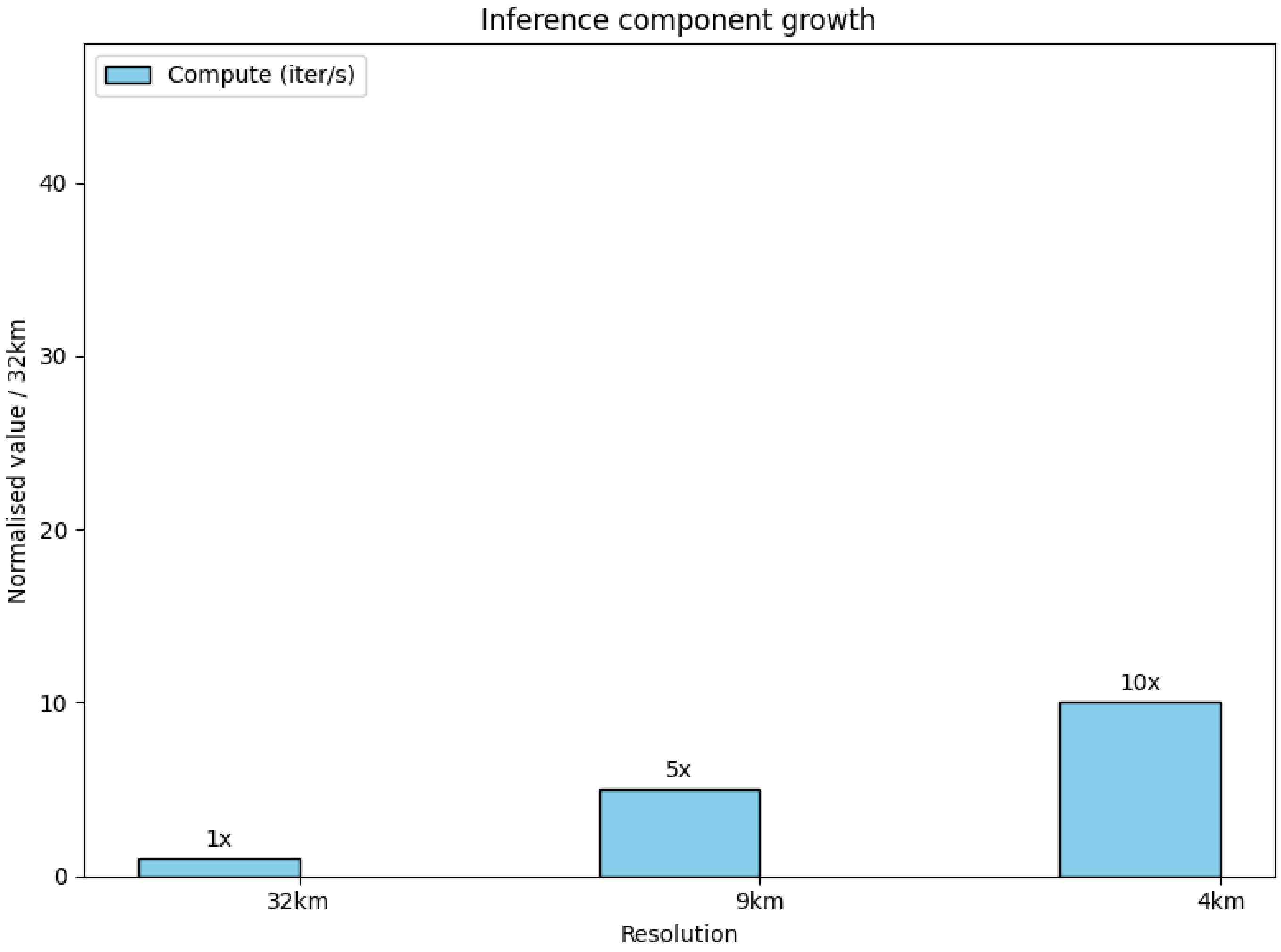
o1280-ds0=/home/mlx/ai-ml/datasets/aifs-od-an-oper-0001-mars-o1280-2023-2023-6h-v1-one-month.zarr
o1280-ds1=/ec/res4/scratch/naco/aifs/inputs/custom/aifs-od-an-oper-0001-mars-o1280-2023-2023-6h-v1-one-month-4gridchunks.zarr
o1280-ds2=/ec/res4/scratch/naco/aifs/inputs/custom/aifs-od-an-oper-0001-mars-o1280-2023-2023-6h-v1-one-month-8gridchunks.zarr
o1280-ds3=/ec/res4/scratch/naco/aifs/inputs/custom/aifs-od-an-oper-0001-mars-o1280-2023-2023-6h-v1-one-month-16gridchunks.zarr
o1280-ds4=/ec/res4/scratch/naco/aifs/inputs/custom/aifs-od-an-oper-0001-mars-o1280-2023-2023-6h-v1-one-month-32gridchunks.zarr

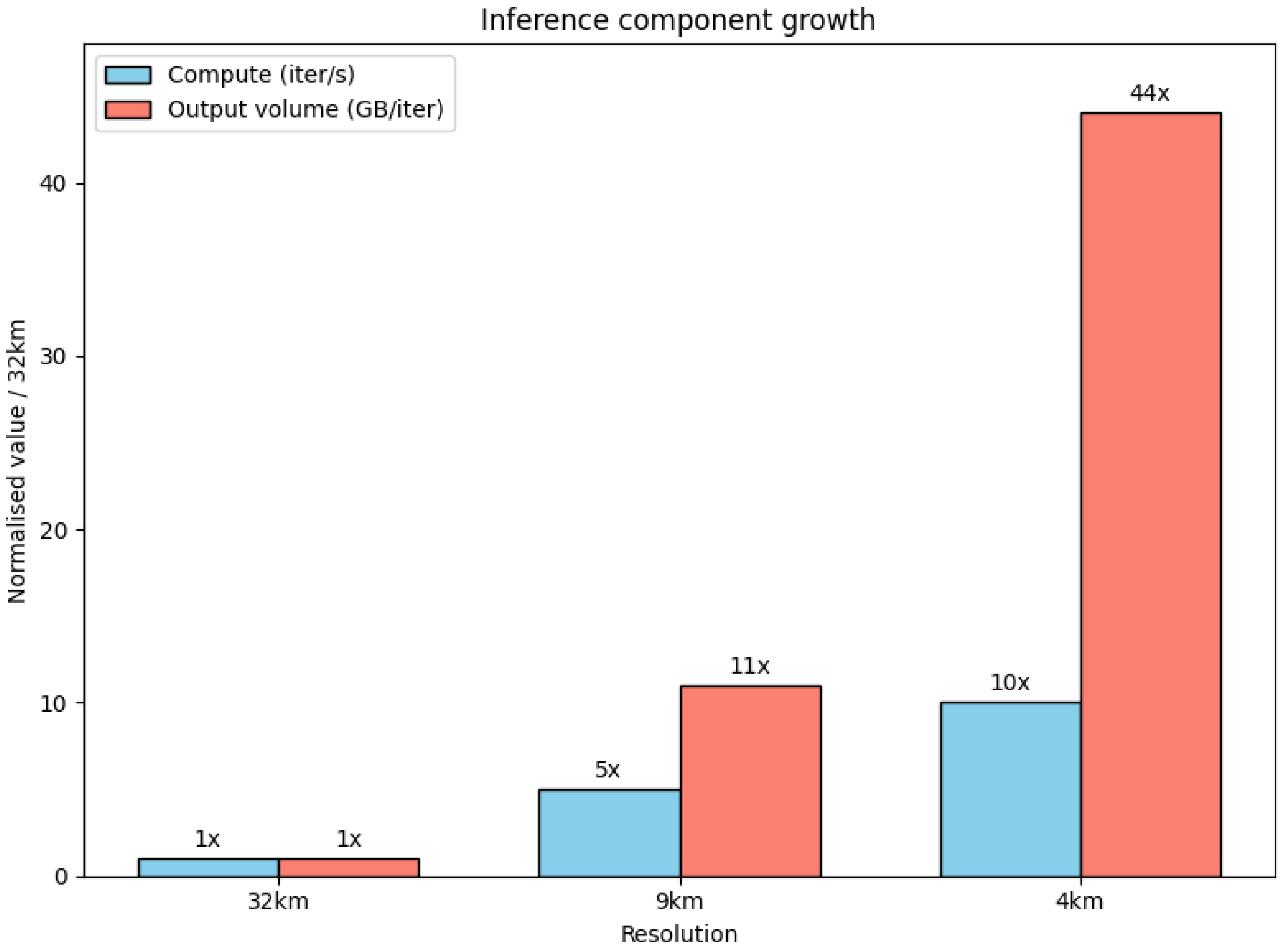
Anemoi inference



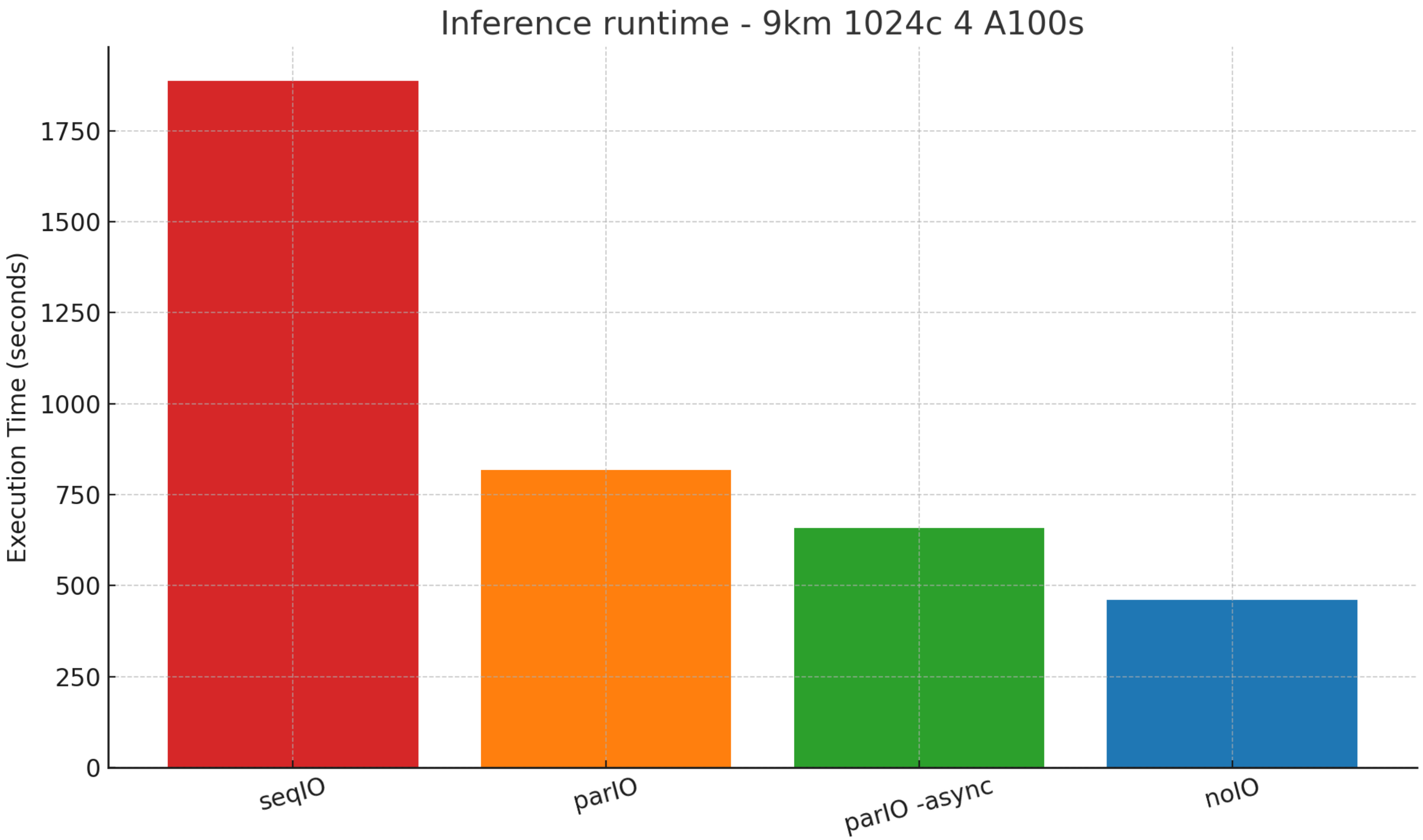
Anemoi inference



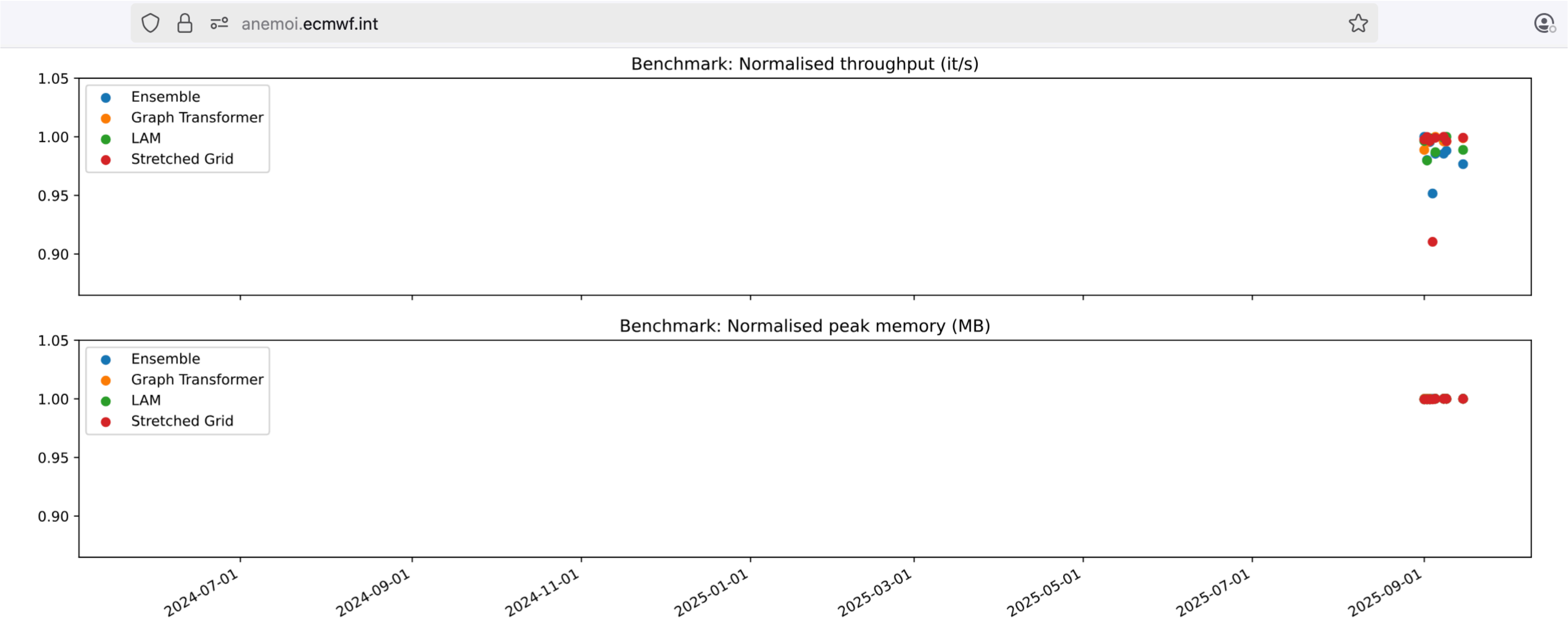

Inference component growth
Compute (iter/s)
Normalised value / 32km
0
10
20
30
40
1x
5x
10x
32km
9km
4km
Resolution


Inference component growth
Compute (iter/s)
Output volume (GB/iter)
Normalised value / 32km
0
10
20
30
40
1x
1x
5x
11x
10x
44x
32km
9km
4km
Resolution

Parallel output

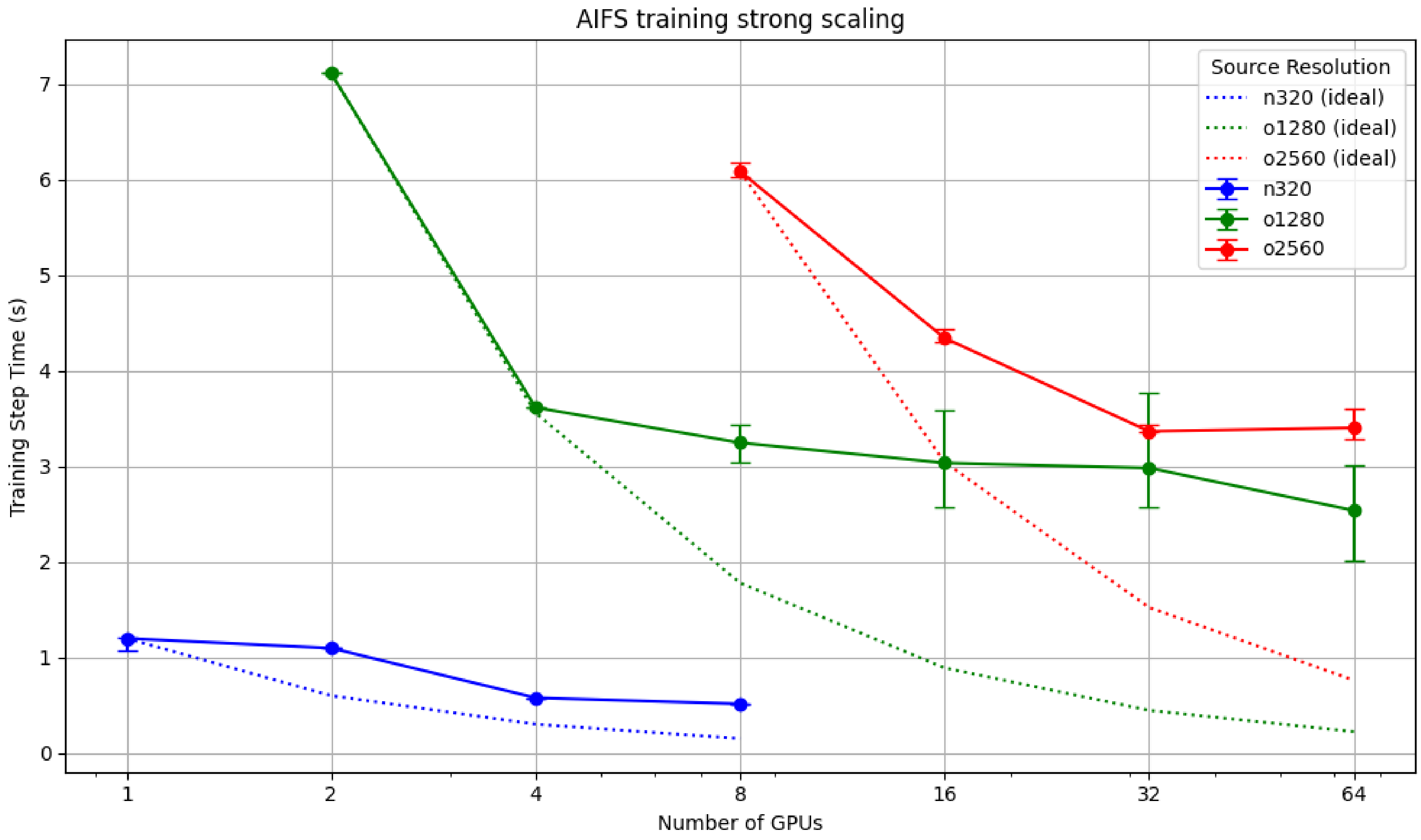


Performance CI testing



- Runs nightly
  - A variety of use cases tested: Global, LAM, Stretched grid and ensemble
  - Throughput and memory usage monitored
  - Model split over 2 GPUs

Next steps: pushing model sharding



Thank you

FMI
METEO FRANCE
À VOS CÔTÉS, DANS UN CLIMAT QUI CHANGE
AEMet
Agencia Estatal de Meteorología
Met Office
Koninklijk Nederlands Meteorologisch Instituut
Ministerie van Infrastructuur en Waterstaat
Norwegian Meteorological Institute
SMHI
Swedish Meteorological and Hydrological Institute
anemoi
ECMWF
MeteoSwiss
DWD
RMI
GeoSphere Austria
The Danish Meteorological Institute

Open-source development driven by
European Meteorological Centers and ECMWF